REGION-BASED DYNAMIC WEIGHTING PROBABILISTIC GEOCODING

A Thesis

by

ZHONGXIA LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Daniel W. Goldberg
Committee Members,	Anthony M. Filippi
	Raghavan Srinivasan
Head of Department,	Vatche P. Tchakerian

August 2014

Major Subject: Geography

Copyright 2014 Zhongxia Li

ABSTRACT

Geocoding has been a widely used technology in daily life and scientific research for at least four decades. Especially in scientific research, geocoding has been used as a generator of spatial data for further analysis. These uses have made it extremely important that geocoding results be as accurate as possible. Existing global-weighting approaches to geocoding assume spatial stationarity of addressing systems and address data characteristic distributions across space, resulting in heuristics and approaches that apply global parameters to produce geocodes for addresses in all regions. However, different regions in the United States (US) have different values and densities of address attributes, which increases the error of standard algorithms that assume global parameters and calculation weights. Region-based dynamic weighting can be used in probabilistic geocoding approaches to stabilize and reduce incorrect match probability assignments that are due to place-specific naming conventions which vary region-to-region across the US. This study tested the spatial accuracy and time efficiency of a region-based dynamic weighting probabilistic geocoding system, as compared to a set of manually corrected geocoding results within Los Angeles City. The results of this study show that the region-based dynamic weighting probabilistic method improves the spatial accuracy of geocoding results and has a moderate influence on the time efficiency of the geocoding system.

DEDICATION

To Dr. Daniel W. Goldberg for his inspiration, support and patience

To my family for their unconditional love

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Daniel W. Goldberg for all the help he gave in these past two years. I would also like to thank my committee members, Dr. Anthony M. Filippi, Dr. Raghavan Srinivasan, and Dr. Tracy A. Hammond, for their guidance and support throughout this research.

I would also like to thank all my friends and co-workers Michael Schwind, Kelsi Davis, Payton Baldridge, and Arron Harmon for their help with my English. Thanks to all my colleagues and the department faculty and staff for making my time at Texas A&M University a great experience. Finally, thanks to my parents for their encouragement and to my girlfriend for her company and love.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF FIGURES vii

LIST OF TABLES ix

1. INTRODUCTION 1

    1.1 Geocoding and geocoding systems 1
    1.2 Motivation and problem statement 5
    1.3 Thesis statement 9
    1.4 Contributions of the research 9
    1.5 Outline of the dissertation 10

2. LITERATURE REVIEW 11

    2.1 Concepts of geocoding 11
        2.1.1 What is geocoding 11
        2.1.2 Input and output formats 12
        2.1.3 Reference dataset 14
        2.1.4 Matching engine (algorithm) 16
        2.1.5 Soundex encoding 19
    2.2 Applications of geocoding 20
        2.2.1 Geocoding in health research 20
        2.2.2 Geocoding in crime research 22
        2.2.3 Geocoding in transportation research 23
        2.2.4 Geocoding in population research 24
        2.2.5 Geocoding in people's lives 24

3. EXPERIMENTAL DESIGN 26

    3.1 Geocoding evaluation metrics 26
        3.1.1 Geocoding spatial error measurement 27
        3.1.2 Geocoding error classification based on weighting score 28

3.1.3 Improvement index 29
3.2 Test data preparation and global-weighting probabilistic geocoding 29
3.2.1 Research area 29
3.2.2 Texas A&M GeoServices website 31
3.3 Manual geocoding correction 32
3.4 Region-based dynamic weighting probabilistic geocoding system design 35
3.4.1 Reference database preparation 37
3.4.2 Dynamic weighting regions generation 42
3.4.2.1 Street signature for ZIP code areas 43
3.4.2.2 Dynamic weighting regions consolidation 45
3.4.2.3 Region statistics database table 45
3.4.3 Weighting calculation 46
3.4.4 Feature scoring 48

4. RESULTS AND ANALYSIS 49

4.1 Results of manual geocoding correction 49
4.2 Result of signature and region generation 50
4.3 Experimental result 54
4.3.1 Results in spatial accuracy 54
4.3.2 Results in time efficiency 58
4.4 Discussion 59

5. CONCLUSION AND FUTURE WORK 61

REFERENCES 63

LIST OF FIGURES

Page

Figure 1.1 Basic components of a geocoding system 2

Figure 1.2 Original map made by John Snow in 1854 4

Figure 2.1 The components of a postal address 13

Figure 2.2 The basic address interpolation workflow 15

Figure 2.3 The matching engine workflow 16

Figure 2.4 An example of a standard weight set. 19

Figure 3.1 The geocoding error 27

Figure 3.2 The research area and the Los Angeles City 30

Figure 3.3 The website interface of Texas A&M GeoServices 31

Figure 3.4 The interface of the manual geocoding correction platform 33

Figure 3.5 The geocoding correction notes 34

Figure 3.6 The region-based dynamic weighting probabilistic geocoding workflow 36

Figure 3.7 NAVTEQ Street Segments Database's folders for each state 38

Figure 3.8 NAVTEQ Street Segments Database's shapefile for one state 38

Figure 3.9 The NAVTEQ SQL Importer 39

Figure 3.10 The results of addresses left-right separation 40

Figure 3.11 Samples of Soundex code fields and their original data fields 41

Figure 3.12 The pre-generated indexes examples for each table 42

Figure 3.13 The examples of street signature set for ZIP code area 44

Figure 3.14 The examples of statistics for region 46

Figure 3.15 The workflow of feature scoring 48

Figure 4.1 The error counting histogram 49

Figure 4.2 Histograms for street signature parameters 50

Figure 4.3 Histograms for street signature parameters 51

Figure 4.4 The region maps (2, 3, and 4 classes) and Los Angeles Communities map ... 53

Figure 4.5 The distribution of improvements index 54

Figure 4.6 The counts of weighting score changes 55

Figure 4.7 Error counts of the global-weighting probabilistic geocoding system 56

Figure 4.8 The influence of the weighting changes 57

LIST OF TABLES

Page

Table 1.1 Examples of each type of error that may exist in input addresses 6

Table 3.1 Definitions of each street signature component 44

1. INTRODUCTION

1.1 Geocoding and geocoding systems

Each time someone uses a GPS navigation system or a mapping service application such as Google Maps or Apple Maps to find a location of an address, they are relying upon a geocoding system. Furthermore, geocoding can help spatially-based research that relies on mapping data which contain house addresses for further spatial analysis. Examples include research in the fields of both health and human geography (Amram et al., 2011; Baker et al. 2012; Balmes et al., 2008; Rushton et al., 2006).

Geocoding, the process of generating a geographic coordinate (often expressed as latitude and longitude) from geographic data, such as street address and ZIP code, has been widely used in daily life and scientific research for at least the last four decades (Dueker 1974; Goldberg 2008; Zandbergen 2008). An example of this process is translating the input address '100 main St Los Angeles, CA 90003' into the latitude/longitude pair '34.051127, -118.243606'. By translating textually formatted location information (addresses) into numeric geographic coordinates, geocoding spatially-enables these data, making them fit for further spatially-based analyses such as mapping, visualization, and spatio-temporal trend modeling (Goldberg et al. 2007; Krieger et al. 2002).

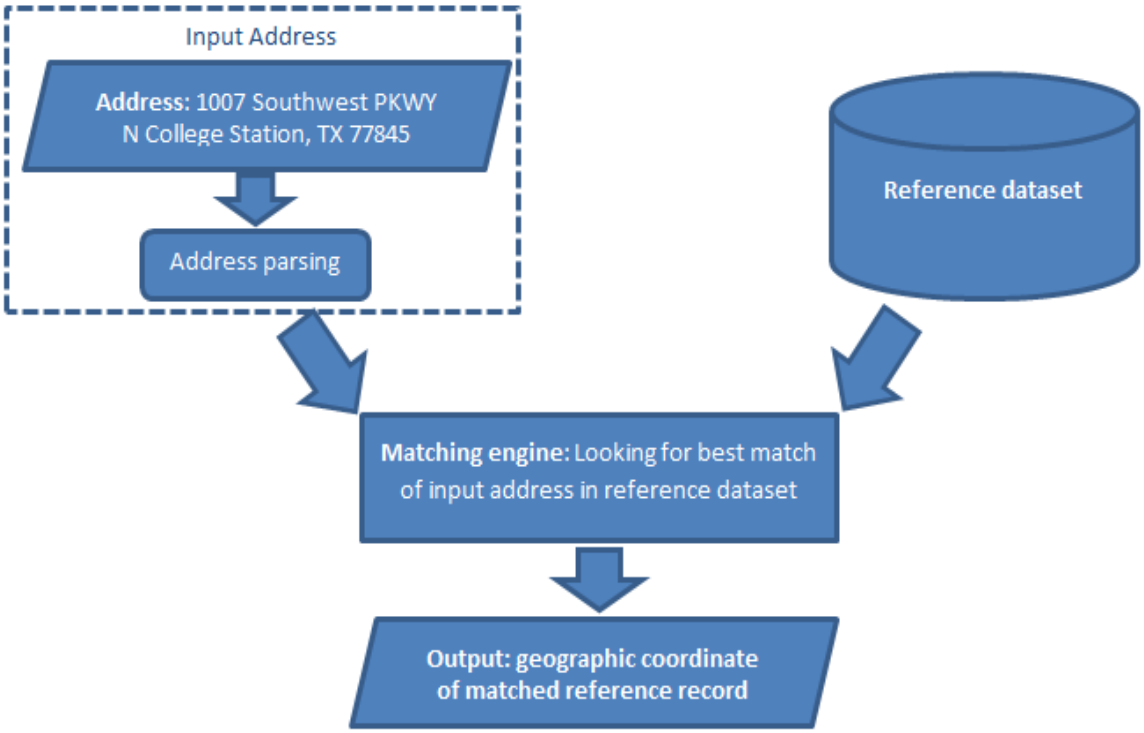


Figure 1.1 Basic components of a geocoding system

A geocoding system, as its name implies, means a system that provides geocoding services. These systems are found as desktop software, web applications, application programming interfaces (APIs), or mobile device software capable of fitting the requirements of a variety of applications. As Figure 1.1 demonstrates, a geocoding system is composed of input addresses, reference datasets, a matching engine, and output

coordinates (Levine et al. 1998; Yang et al. 2004). The matching engine is designed to search through the reference dataset in order to find the best matching record compared to an input address and compute and return the geographic coordinate of this record as the geocoding result.

Researchers have advocated using geocoded data in geographic information systems (GISystems) as an essential component of spatial analysis. Geocoding been in use for more than four decades, and has been widely exploited across many research disciplines and application domains. These include epidemiologic research (Bonner et al. 2003; Oliver et al. 2005; Schootman et al. 2007; Wheeler et al. 2012; Zhan et al. 2006); cancer research (Goldberg et al. 2012; Krieger et al. 2002; Krieger et al. 2001; Rushton et al. 2006; Rushton et al. 2010); transportation analysis (Dueker 1974; Ozimek et al. 2011; Qin et al. 2013); crime analysis (Mamalian et al. 1999; Police Foundation 2000; Ratcliffe 2004); and population studies (Chen et al. 1998; Gilboa et al. 2006; McElroy et al. 2003; Robinson et al. 2010).

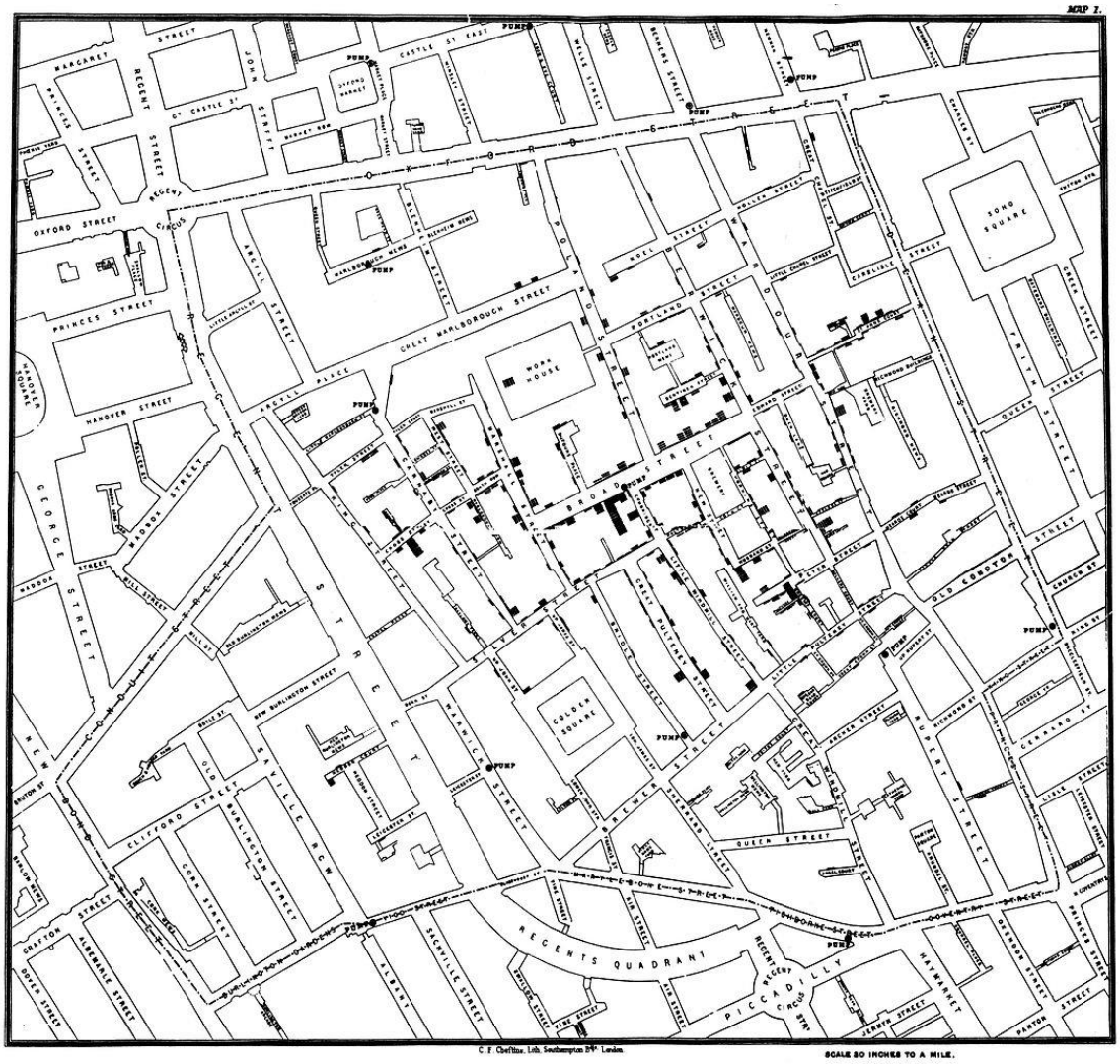


Figure 1.2 Original map made by John Snow in 1854

One of the classic applications in GIS, Dr. Snow's dot map (also known as the Cholera map or the ghost map) can be seen as a potential exemplar use of a geocoding system. John Snow (1813 –1858) was an English doctor whose story is often taught in Geography, GIS, Spatial Statistics, and Epidemiology courses. His study is displayed in Figure 1.2. Dr. Snow used a dot map to display the locations of cholera death cases occurring in London in 1854. The key insight of his work is that through the visualization of these incidences as symbols

on a map, it became apparent that there were more cases clustered near a particular water well which led to the discovery that the well was contaminated and the source of the disease outbreak. This study marked the establishment of the modern-day epidemiology.

In Dr. Snow's study, the fundamental step of mapping deaths due to cholera cases could have been accomplished by geocoding the addresses of the death cases and mapping them with a GISystem, had geocoding systems been available at the time. Instead, mapping and drawing of all cases was done manually, one by one onto a map. Geocoding would have reduced the time and effort necessary to accomplish this same task, perhaps leading to timelier conclusions and fewer deaths.

1.2 Motivation and problem statement

International utilization within many academic, government, and business fields has made it extremely important to that geocoding results be as accurate as possible. Geocoding acts as the fundamental spatial data generator in many research disciplines and application domains (Bell et al. 2006; Bonner et al. 2003; Costello et al.2009; Dueker 1974; Oliver et al. 2005; Vine et al. 1997; Zhan et al. 2006). As such geocoding systems have become a crucial procedure necessary to make research progress in many spatially-based fields. In many spatially-based studies that investigate the locations of people, geocoding is the first process undertaken after data collection. This makes geocoding an important factor in the accuracy of subsequent research methods which utilize these data and the results they produce. Any errors that are generated during the geocoding process may be magnified by further spatial processing or analyses, which may lead to essential differences in study outcomes. Therefore, reducing the error during the geocoding progress has been noted as pressing need in the literature (Goldberg et al. 2012).

Table 1.1 Examples of each type of error that may exist in input addresses

	Number	Pre-directional	Street name	Suffix	Post-directional	City	State	ZIP code
Correct address	800	N	Main	St		Houston	TX	77002
Incorrect number	**801**	N	Main	St		Houston	TX	77002
Missing part	800		Main	St		Houston	TX	77002
Typo	800	N	**Man**	St		Houston	TX	77002
Incorrect directional	800	**S**	Main	St		Houston	TX	77002
Incorrect street type	800	N	Main	**Rd**		Houston	TX	77002
Incorrect region scale	800	N	Main	St		**North Houston**	TX	77002
Incorrect ZIP code	800	N	Main	St		Houston	TX	**77017**

Complications that lead to inaccurate geocoding include input address uncertainty errors and missing input address missing attribute problems that are contained within the input datasets processed by geocoding systems. Incomplete or incorrect input address data will cause inaccuracy and uncertainty in resulting geocoded data. Errors like missing address components, spelling mistakes (typos), incorrect address directionals, street types, sub-region, and city names, and incorrect ZIP codes can and are generated during data collection process and affect the quality of geocoded data. Each of these errors results in different amounts of spatial displacement error in output geocoded results. Table 1.1 gives examples of each type of error that may exist in input addresses.

Many traditional geocoding methods assume that input addresses are correct and well formatted. Without taking the possible errors previously mentioned into account (in

addition to many more), standard geocoding methods may suffer serious problems when geocoding large databases of address records. This is a particularly pressing issue when determining a precise location is critical to health, safety, property, or policy efforts. Probabilistic geocoding has been developed as one approach to assist in the selection of the best matching results between an input dataset and large reference databases, and has been implemented in several geocoding systems (O'Reagan 1987; Jaro 1984) to improve their performance.

Probabilistic matching approaches compare uncertain input addresses using an approach that separates an input address into its multiple address components and associates a weight, or importance, to each component based on the density of values as observed within a reference database. The weight for each field is determined by the probability of the input field matching with the field of one record in the reference database that is a true match and the probability of the input field matching with one random record from the reference database. The totals of the weights for all the components of postal address are usually represented by a number between 0 and 100 (Francis P. Boscoe, 2008). Traditionally, each weight is calculated based on the density of each possible value of each of the street address attributes and the same set of weights is applied nationwide. In the current study, this approach which uses a globally-defined set of attribute weights uniformly across all input data, regardless of the location of the input address data, is termed the *global-weighting* approach.

In the US, as in other places around the globe, addressing systems vary by region given the history and policies of particular places. The ways that streets are named, the prevalence of numeric street names or single-letter alphabetic streets names, and the pattern

and distribution of street segments and house numbers are all characteristics that reflect street naming and house numbering conventions which, in the US, are controlled at the local government or regional level. As one moves from region to region in the US, one observes different values and densities of address street name attributes, due to the non-stationarity of these naming conventions across the US as a whole. Different regions in the US commonly prefer different conventions for naming streets and indicating street types. For example, sections of Los Angeles City frequently contain Spanish names (e.g., 'Los Feliz Blvd' and 'La Paz Dr' in ZIP code 90027), while most of the streets in Bellevue, Washington State are named using numerical street names with either pre-directional or post-directional (or both) (e.g., '140th Ave SE', 'NE 8th St', and 'West Lake Sammamish Pkwy NE'), and New York City has a preponderance of streets with numerical names (e.g., 'W 57th St' and '5th Ave', in Manhattan).

This heterogeneity of addressing and naming systems increases the error of global-weighting geocoding algorithms that employ the same probability weights to all input addresses regardless of region. The application of the same set of weights for the entire country results in different levels of accuracy by region since the distribution of types of address attributes are not the same everywhere (non-stationarity). The globally-weighted approach which assumes constant weights globally fails to account for the local and regional differences in values and densities of attribute, meaning that the characteristics of a *place* are not leveraged to increase the accuracy of geocoding for that particular location. As a result, these methods produce lower quality data than could be produced using locally relevant modeling techniques. The central thesis of the following work is that incorporating local street naming conventions and address distribution characteristics will improve

results based on local conditions present in the region the input address is located. The proposed method, termed *region-based dynamic weighting probabilistic geocoding* will provide more accurate probabilistic geocoding result as a better data source for modern GISystems.

1.3 Thesis statement

Region-based dynamic weighting can be used in probabilistic geocoding approaches to stabilize and reduce incorrect match probability assignments which are due to place-specific naming conventions which vary region-to-region across the US.

1.4 Contributions of the research

This study contributes to the knowledge in the area of geocoding in general, but more specifically, to the research, development, and application of a novel geocoding technique which uses dynamic region zoning and dynamic attribute weighting to improve the spatial results of geocoding systems. To achieve this goal, spatial clustering and spatial index technologies are developed to compute regional differentiation of reference data. Also developed is an automated method which breaks the world up into a series of contiguous regions which share addressing characteristics. In order to optimize existing probabilistic geocoding methods and improve resulting spatial accuracy, a probabilistic geocoding system was built to test the proposed method which utilizes differences in cultural and street naming conventions across disparate regions.

This research contributes to society because, fundamentally, it improves the accuracy of geocoded results. This progress will make the geocoding process more reliable and of increased value to researchers, agencies, commercial groups, and individuals due to the fact that an incorrectly formatted or wrong address will still result in an accurate match.

1.5 Outline of the dissertation

The remainder of this thesis is organized as follows.

Chapter 2 presents a detailed literature review on the current state and utilization of geocoding techniques which are used as the motivation for and basis of the current research.

Chapter 3 describes the details of the research experiment, including the data preparation, system design, and development.

Chapter 4 outlines the results of the experimental design and provides a discussion of the results through several in-depth analyses of notable findings.

Chapter 5 concludes the thesis by emphasizing major achievements and describing the potential for future work.

2. LITERATURE REVIEW

2.1 Concepts of geocoding

There are many different ways to describe locations on the surface of the Earth. These include for example, place names (e.g., 'New York Central Park'), postal addresses (e.g., '100 Main St Los Angeles, CA 90002'), relative directions (e.g., 'one mile south to the museum'), and geographic coordinates (e.g., latitude and longitude: 30.614910, -96.342295). Postal addresses have been, without a doubt, the most commonly used locational format in research studies because they are widely used, well formatted, and easy to remember. However, these textual addresses cannot be recognized and utilized by a GISystem as easily as the digital coordinates (Goldberg 2008). With the development of GIS, how to translate a descriptive sentence such as '800 main St N' into a digital coordinate that could be understood by a computer system (GISystem) has become an essential topic for GIScience.

2.1.1 What is geocoding

Geocoding and reverse geocoding systems appear as the translators between descriptive human language and numerical geographic coordinates. Geocoding means the process of generating a geographic coordinate (often expressed as latitude and longitude) from descriptive geographic address or location information. Conversely, reverse geocoding describes the process of converting digital geographic coordinates (such as latitude/longitude data obtained from global positioning system [GPS] devices) into human language (such as street addresses).

Because addresses are fuzzy descriptive data which must be converted into digital coordinates, a great number of the prior studies into geocoding processes have engaged in reducing spatial error and uncertainty in geocoded results (Goldberg et al. 2010; Goldberg et al. 2012; Krieger et al. 2001). The study presented within this thesis also focuses on improving the accuracy of geocoding systems. In contrast, most of the research tackling reverse geocoding has been devoted to improving the query process time efficiency (Zarem et al. 2006).

As mentioned in **Chapter 1**, a geocoding system (geocoder) is primarily composed of one or more input addresses, one or more reference datasets, a matching engine, and one or more output coordinates (Levine et al. 1998; Yang et al. 2004). The following sections describe each of these components in detail.

2.1.2 Input and output formats

The input street address used as the input data given to a geocoding system is the descriptive locational text that represents one geographic point on the surface of the Earth. The variety inherent in human language leads to a mass of practices used to describe geographic locations. These include descriptors such as postal addresses (Ge 2005; Goldberg et al. 2007; Rushton et al. 2006), street intersections (Guo et al. 2010; Levine et al. 1998), named geographic features (Lee 2002; Davis et al. 2003; Taranenko et al. 2011), ZIP codes (Krieger et al. 2002; Krieger et al. 2003), and free-formatted textual locational descriptions (Wieczorek et al. 2004). Among these different formats, postal addresses are favored and most commonly encountered in datasets processed by geocoding systems, especially those processed by scientific researchers, who obtain these data as part of many data collection processes. In the US, researchers have identified postal addresses as the

most common input to geocoding systems (Ge 2005; Goldberg et al. 2007; Rushton et al. 2006). The focus of the research presented herein limits the scope of potential input data types to postal address data only.

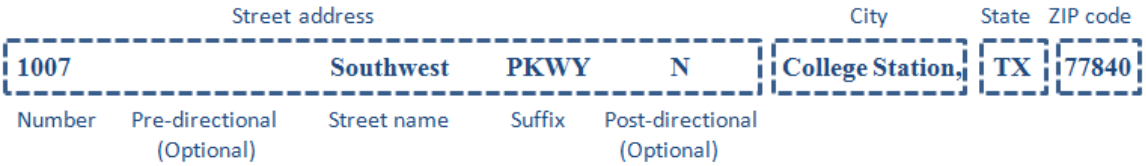


Figure 2.1 The components of a postal address

Figure 2.1, illustrates the components that comprise a typical postal address: Street address, City, State, and ZIP code. Street address is usually formatted by combining a series of fields including the address number, the pre-directional (optional), the street name, the street suffix, and the post-directional (optional). Additionally, some postal address data may contain a suite number and suite type. However, the majority of existing geocoding systems in use today do not take suite level information into account when generating geocode outputs because of accuracy limitations in most reference datasets available to geocoding systems – most reference datasets do not include information down to the suite level.

A geographic coordinate system is a coordinate set generated to indicate every possible location on the surface of the Earth with a set of numbers and/or letters (Burrough et al. 1998). Most geocoding systems output coordinates in the form of latitude and

longitude. This is also the data format of the Global Positioning System (GPS), one of the most popular space-based satellite navigation systems.

2.1.3 Reference dataset

Reference datasets are fundamental to geocoding systems. The accuracy and integrity of a reference dataset is a prerequisite for any geocoding system that is intended to produce high-quality results. Because the construction of a high-quality reference dataset requires intense work and high costs, most of the datasets that are used by geocoding systems are provided by either a national organization or large commercial company. Typically, most of the commonly used reference datasets are classified as two different types: street segment (line vector data) and address point (point vector data), although polygon reference datasets exist and are sometimes used as well.

Street segment reference datasets such as the US Census Bureau's TIGER/Line shapefiles database (U.S. Census Bureau 2013) and the NAVTEQ Street Segments Database (Here 2013) utilize an address interpolation algorithm to generate a position for an input address using the address range associated with the street segment matched to by the geocoding system. Because street segment databases are the type of reference dataset most commonly used in geocoding systems, numerous researchers have attempted to understand and improve the accuracy of geocodes produced using these interpolation methods (Bakshi et al. 2004; Dueker 1974; Nicoara 2005; O'Reagan et al. 1987; Ratcliffe 2001).

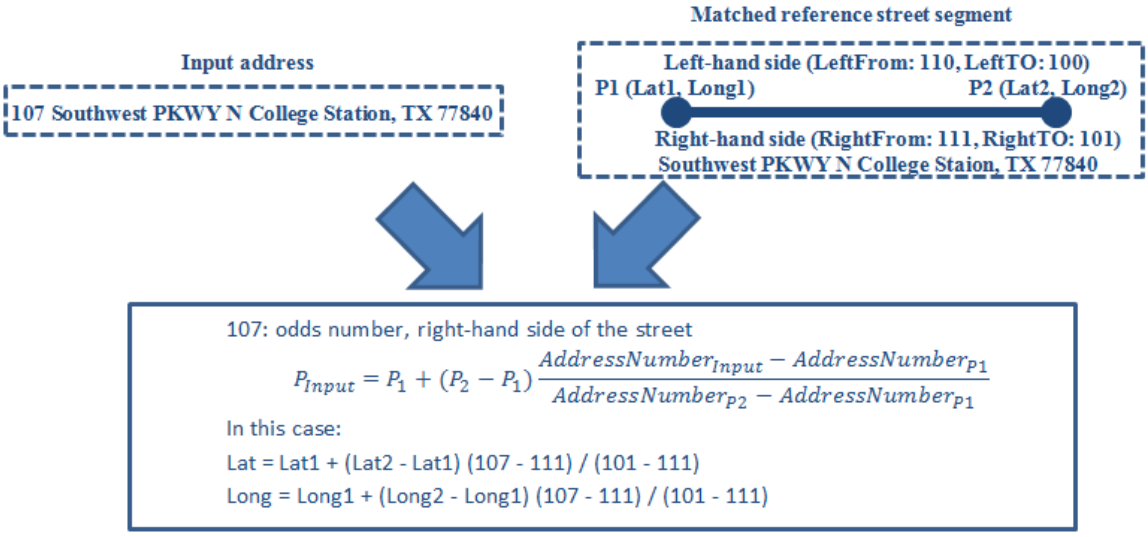


Figure 2.2 The basic address interpolation workflow

$$P_{Input} = P_1 + (P_2 - P_1)\frac{AddressNumber_{Input} - AddressNumber_{P1}}{AddressNumber_{P2} - AddressNumber_{P1}} \qquad (2.1)$$

Figure 2.2 describes a basic workflow of the address interpolation geocoding method. After a best-matching street segment is found by searching a reference dataset, address interpolation and linear interpolation are used to estimate the latitude and longitude to represent the output coordinate for an input address. The first step in this process is to determine whether the input address should be a part of the right-hand or the left-hand sides of the street segment based on the parity of the house number and those associated with each side of the street segment. Based on the address interpolation method defined by Figure 2.2 and Equation 2.1, both latitude and longitude are calculated (Zandbergen 2008).

Address point (parcel centroid point) databases such as Boundary Solutions' National ParcelMap Data Portal (NPDP) and NAVTEQ Address Point Databases have been used as

reference datasets by some geocoding systems. As a database that stores every existing address in the whole country, most of these point reference databases contain millions of records. Due to these large data sizes, these types of databases typically see low query time efficiency meaning that indexing techniques and technologies are needed to assure adequate geocoding query processing time. In the big data age, technologies which have been designed for non-relational (or less relational) big datasets are becoming increasingly common. These include Not Only SQL (NoSQL) databases such as MongoDB, Google BigTable, and Oracle NoSQL Database. These data storage formats have been developed to begin to address query speed issues in massive datasets, such as those used in geocoding systems (Rischpater et al. 2013).

2.1.4 Matching engine (algorithm)

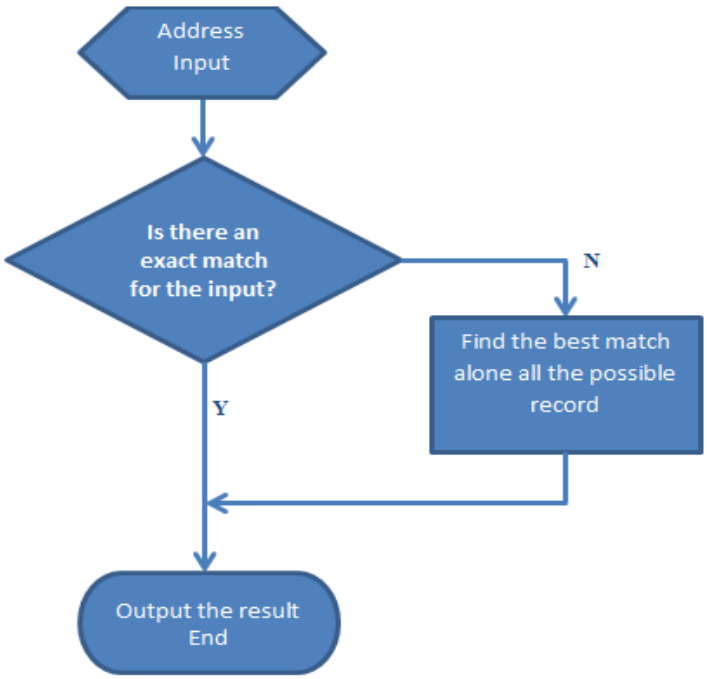


Figure 2.3 The matching engine workflow

The main task of a matching engine is to find the best match (that with the highest likelihood of being correct) for an input address given all available addresses (geographic objects) contained within a reference dataset. Matching processes can be categorized as either deterministic or probabilistic. Figure 2.3 demonstrates a basic workflow of a matching engine. First, the geocoding system searches for an exact match for the requested input address. If an exact match cannot be found, the system will compare all possible matches and identify the most likely match (best match) for the input address. How to select the best match, both efficiently and accurately, is the main question under consideration in the current study.

Probabilistic geocoding systems have been developed to help select the best matching result from reference databases (O'Reagan 1987; Jaro 1984). A probabilistic geocoding system scores all of the possible records available within a reference dataset against an input address based on the attribute similarity and the importance of each of the components that make up an address. A probabilistic matching algorithm accomplishes this by first calculating a weight for each field of the input address representing how important each field of the address is.

$$\text{Weight} = \ln(m/u) \tag{2.2}$$

As Equation 2.2 describes (Boscoe 2008), the weight results from two separate probabilities (*m* and *u*). The first of these, *m*, represents the probability that the input field agrees and the match between an input address field and that of the reference data field is a true match. Since the method is comparing all similar records alone, this value is usually

either very high (close to 1 – 100% probability) or very low (close to 0 – 0% probability). For example, if we consider an example where an input street name is provided as 'Texas' and the reference street name is also 'Texas', the value of the probability (m) should be 1. Even if there are a few minor typos, like 'Texes', it should still be consider a high match (like 0.95). More discussion on this topic is provided in Section 2.1.5 during the description of programmatically recognizing typo errors.

The second probability, *u*, is the probability that the value of the input field agrees with the value the field on a random record in the reference dataset. In general, the value of this probability is low. For example, consider a dataset that includes 100 million street segments, of which 400,000 are named 'Main'. In this case, the probability of a randomly selected record happens to have a name of 'Main' is 0.004. Researcher have found that in practice, the *u* probability of the pre-directional and the post-directional fields (N, S, E, and W) are near 0.25 (Boscoe 2008).

The weight for each field of the input address can be calculated using Equation 2.2 once both *m* and *u* are known. For the street name component of an address, assuming *m* is 0.95 and *u* is 0.004, the weight would be 5.47 (Equation 2.2). For the post-directional field, assuming an *m* of 1 and a *u* of 0.25, the weight would be 1.39. This result makes intuitive sense, because it is represents the fact that a street name should be weighted quite a bit higher than the post-directional field.

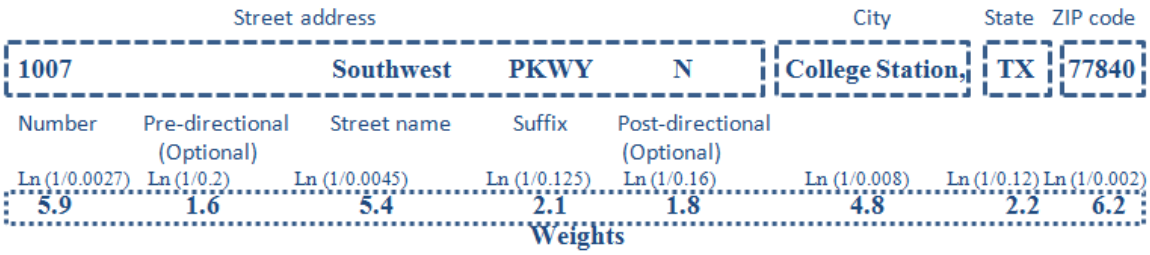


Figure 2.4 An example of a standard weight set.

Traditionally, probabilistic geocoding engines assume *m* is 1 and generate a set of standard weights for each component of the address based on the value of national-level reference datasets. Figure 2.4 gives an example of the standard weight set. More discussion about standard weight sets are given in section 3.4.3.

2.1.5 Soundex encoding

Misspellings (typos) like ‘Mapple’, ‘Univorsity’, and ‘Las Angeles’ for the correct versions ‘Maple’, ‘University’, and ‘Los Angeles’ are commonly generated during a data collecting or data entry process. Soundex encoding, a phonetic algorithm for indexing English words by the way they are pronounced, is designed to programmatically recognize English words despite the minor differences in spelling (Zandbergen 2008). This algorithm has been widely used in geocoding systems in order to match misspelled words in the input data, with correct reference datasets (Boscoe et al. 2002; Goldberg et al. 2007; Yang et al. 2004). For example, the Soundex code for the word ‘University’ is ‘U516’, while the Soundex code of the misspelled word ‘Univarsity’ is ‘U516’ as well. Therefore, to a geocoding engine which utilizes Soundex, these two words would be recognized as a match.

2.2 Applications of geocoding

Geocoding technology plays an important role both in the everyday lives of people and in scientific applications. In this section, the rich literature on geocoding technologies is discussed to demonstrate the utility and importance of these systems.

2.2.1 Geocoding in health research

Health data processing was one of the first applications of geocoding systems (Rushton et al., 2006). Research in this area has studied the relationships between the location of cancer and other epidemic diseases and environmental factors. For example, Amram et al. (2011) analyzed the impacts of air and noise pollution on children's health development by exploring the relationship between a child's school's location and the child's health condition. To accomplish this, the authors investigated the distance of schools to major roads in Canadian cities using geocoding technology. The authors found that traffic-generated air and noise pollution have serious effects on children's health development. Since children spend most of their time at school, the location of the school may be an important factor of epidemiologic exposure. Their results indicated that a large amount of students at public schools in Canada faced high levels of air and noise pollution when they were at school in low income area. The locations of schools were shown to have potential negative impacts on students' health development.

Similarly, researchers have found that exposure to traffic can cause asthma in children using geocoding systems. Balmes et al. (2008) conducted a study to analyze the relationship between traffic exposures and health status in adults. Geocoding technology was used to generate the detailed roadway and address information that powered the study's analyses. The result indicated that traffic exposures can decrease adults' lung

functionality because traffic exposures increase the morbidity of asthma. The study found that exposure to high density traffic and distances to the nearest roadway have strong impacts on lung function in adults.

There are also studies which examine the relationship between cancer incidences and the socioeconomic characteristics of people. For instance, Krieger et al. (2005) conducted a survey to test the links between US socioeconomic gradient and breast cancer. Geocoding technology was used to generate the breast cancer incidence points from address data. These geocoded cancer incidences were then linked to a socioeconomic gradient. This seminal work found that the cancer incidences vary by race or ethnicity.

Other studies have investigated the relationship between health condition and health service accessibility. Continelli et al. (2010) analyzed the relationship between local doctor supply, the possibility of having a primary care doctors, and the possibility of receiving preventive health examination. The results indicated that the local doctor supply has an impact on the possibility of having a primary care doctors and affects the preventive service which indicates that the health service accessibility has an influence on health development. Ngamini Ngui et al. (2011) analyzed the spatial accessibility of mental health service in Canada. Through the analysis of the potential demands for mental health service and the supply of mental health services, the results indicated that the mental health services are unequally distributed in the southwest of Montreal. This research provides an indicator of the need for the improvement in the distribution of health service

A second trend in geocoding research relates to geocoding accuracy in health research. Goldberg et al. (2012) investigated the effects of interpolation method on county-level cancer rates when case information is geocoded to the ZIP code level. Schootman et al.

(2006) examined the spatial accuracy and geographic errors of four geocoding algorithms within the context of epidemiologic research. In this work, the 'point-in polygon' method and the 'look-up-table' algorithm were compared to intersect addresses into census area. Zimmerman et al. (2010) quantified the effects of local street network conditions on the spatial accuracy of batch geocoding with epidemiologic cases.

In sum, geocoding technology plays a significant role in the development of spatially-based health research. It provides important evidence that improves the application of GIS within health research.

2.2.2 Geocoding in crime research

Geocoding technology has also been used widely in crime research, especially during early research developments in the field. Geocoding contributed to better ways of understanding the relationships between crime, socioeconomic, demographic, and geographic factors. Geocoding provided a method that police agencies could use to ultimately reduce the incidence of crime.

Ceccato et al. (2004) compared crime patterns at two points in time to analyze the relationship between crime occurrence and investments in transportation services in a border area. The results indicated that improved transportation systems generated an increase in mobility, but the total number of crime did not increase. However, since the improved transportation systems that cross 'open' borders created easier access to places, this led to changes in smuggling routes and facilitated human trafficking. Andresen (2006) investigated the spatial autocorrelation between local crime rates and socioeconomic features at the census area level. The results indicated that high unemployment and the presence of young people had a very strong relationship with local criminal rates.

Gruenewald et al. (2006) analyzed the relationship between alcohol sales volume in time and local violence rates. The study indicated that greater percentages of minorities and lower median household incomes increased the rates of violence. The greater number of alcohol outlets also increased the violence rates. Each of the above studies utilized geocoding as a fundamental component of the research design.

2.2.3 Geocoding in transportation research

The application of geocoding technology in transportation research has been primarily applied in two contexts: improving the condition of transportation systems and decreasing traffic crashes.

Chou (1995) developed a decision support system for public bus routing, route order mapping, and passenger addresses geocoding. This system integrated GISystem mapping functionalities and geocoding methods with other technologies. Overall, six systems were contained in the complete system, including a user-based routing function which allowed the user to select optimal routes, a walking-distance calculation function which identified street addresses that were within a user-specified walking distance from users' initial GPS location, a bus-stop component which generated optimal bus stops according to travel demand, a passenger plotting module which geocoded passengers' addresses, plotted passenger's location, a user-based routing function, and a walking-distance calculating function. The authors' results showed that the system improved the condition of urban transportation systems.

Geocoding is an important technology for analyzing the relationship between vehicle collisions and transportation system conditions which can provide invaluable resources for injury prevention researchers. Park et al. (2011) applied a post mile referencing dataset

within a geocoding system in order to geocode collisions on expressways in South Korea and identified the most appropriate methodology for Korean expressways in particular. The results indicated that the geocoded database of expressway collisions improved the traffic safety and reduced fatalities. Qin et al. (2013) developed a 'Crash-Mapping Tool' to geocode locations of police crash reports and create pinpoint maps for all crashes. This integrated crash map provided an effective method for crash analysts to locate where crashes happened on the highway.

2.2.4 Geocoding in population research

Geocoding technology has been used in population estimation. Geocoded address data and housing-unit methods are often used to estimate small-area populations. However, the incompleteness of georeferenced address-based datasets has been known to cause low accuracy in population estimation. Baker et al. (2012) evaluated the influence of incorrect geocoding on accuracy in small-region's population estimates. The study indicated that incomplete geocoding potentially introduced large amounts of error in population estimates.

2.2.5 Geocoding in people's lives

Geocoding technology is also an important component in many branches of people's lives. For example, it is used when developing a spatio-temporal method for activity location reconstruction. Individual-level travel survey datasets are a valuable resource for analyzing human movement. However, quality issues within travel survey data have limited the effectiveness of these data depending on how geocoding is accomplished. Horner et al. (2012) presented a method to geo-enable activity locations from travel surveys that could not be accurately geocoded. The proposed method estimated the probabilistic

locations of missing trip stop points. The method generated estimated locations for unreferenced destinations, which improved the usefulness of the survey data.

Geocoding technology has also been used in the assessment of environmental impacts on residential property prices. Kim et al. (2013) analyzed impact of light rail on residential property prices. Chasco et al. (2012) analyzed the impact of noise and air quality on house prices.

Other studies have used geocoding technology to measure alcohol outlet density, analyze the impact of tobacco sale volume, and estimate the results of smoking cessation efforts (Matthews et al. 2011, Han et al. 2014).

3. EXPERIMENTAL DESIGN

The experiment described here was designed to implement the region-based dynamic weighting probabilistic geocoding system and evaluate it. First, test addresses were geocoded by a global-weighting probabilistic geocoding system. Second, the geocoded test data were manually corrected with a manual geocoding correction platform. Third, the test input addresses were geocoded by the region-based dynamic weighting probabilistic geocoding system. Finally, the geocoded results from the global-weighting geocoding system and the region-based dynamic weighting probabilistic geocoding system were compared with the manually corrected results to evaluate improvements in spatial accuracy resulting from the proposed method.

3.1 Geocoding evaluation metrics

The experiments described herein were designed to facilitate the evaluation of the region-based dynamic weighting approach described in two ways. The first was to assess improvements in spatial accuracy, meaning that geocodes produced with the proposed method would be closer to ground truth values than those produce using global-weighting (non-dynamic) methods. The second was to assess the representative accuracy in match scores that were assigned by the region-based dynamic weighting and global-weighting methods. The hypothesis tested in this work was that the proposed method should have generated different match scores than the global-weighting method. Improvements in match scores would mean that results which were previously false positives (those scored erroneously as matches when they should not have been) should receive lower scores using the region-based dynamic weighting approach. Similarly, previous false negatives (those

scored erroneously as non-matches when they should have been matches) should receive higher scores using the region-based dynamic weighting approach.

3.1.1 Geocoding spatial error measurement

This study assumes that the coordinates derived from the manual geocoding correction progress are true points following the approach described in Goldberg et al. (2008). As such, geocoding error is defined by the bias (distance) from the geocoded result point to the real point (corrected longitude and latitude).

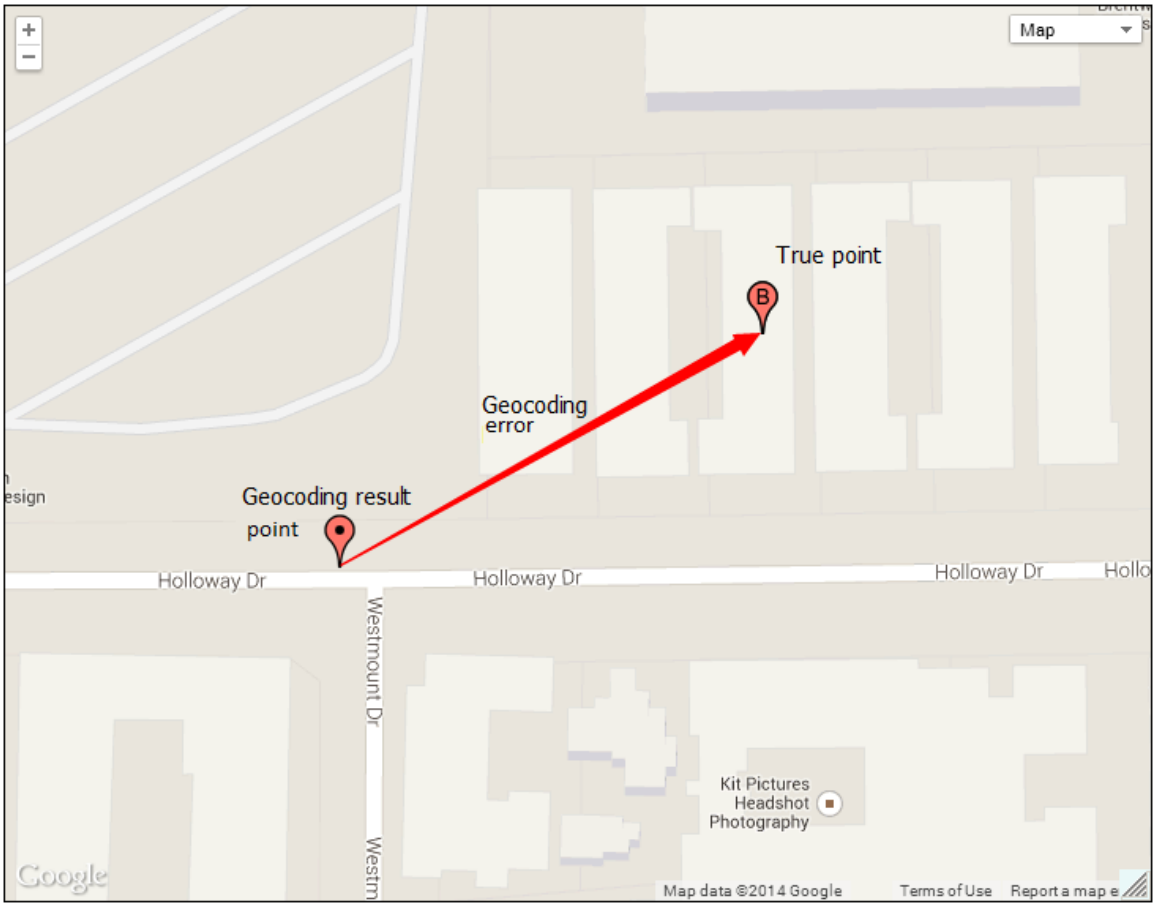


Figure 3.1 The geocoding error

Figure 3.1 demonstrates an example of geocoding error. Geocoding error in this case means that the geocoded point and the true point are not in the same location. These errors, or shifts, from the true point to the geocoding result point can be represented by either vector or distance metrics. This study focused on the distance from the geocoding result point to the true point.

Great Circle Distance was used as the distance calculation algorithm:

$$D_{Lon} = Lon_1 - Lon_2 \tag{3.1}$$

$$D_{Lat} = Lat_1 - Lat_2 \tag{3.2}$$

$$a = \sin(d_{Lat}/2)^2 + \cos(Lat_1) * \cos(Lat_2) * \sin(d_{Lon}/2)^2 \tag{3.3}$$

$$c = 2 * atan2\left(\sqrt{a}, \sqrt{1-a}\right) \tag{3.4}$$

$$Distance = R * c \quad (R \ \ is\ the\ radius\ of\ the\ Earth: 6378.1km\ or\ 3961.3miles) \tag{3.5}$$

3.1.2 Geocoding error classification based on weighting score

Geocoding errors caused by incorrect weighting scores were classified as two types of errors: false positives and false negatives. False positive errors mean that the input addresses failed to match with the correct locations due to computed weighting scores that were too high. False negative errors means that the input addresses failed to match with the correct locations due to weighting scores which were too low. Generally, false negative errors will return ZIP code centroid geocode quality types while false positive errors will return street address quality locations, but not the correct location.

3.1.3 Improvement index

The research presented here developed the concept of an improvement index which was designed as a tool for detecting the differences resulting from the region-based dynamic weighting method.

$$\text{Index}_{Improvement} = \text{Distance}_{Old} - Distance_{New} \tag{3.6}$$

Equation 3.6 illustrates the definition of the improvement index of geocoding accuracy. This measure is the distance differential between the accuracy of global-weighting probabilistic geocoding system and region-based dynamic weighting probabilistic geocoding system.

3.2 Test data preparation and global-weighting probabilistic geocoding

3.2.1 Research area

Los Angeles City was selected as the research area for these experiments due to the availability of test and reference data and in-depth knowledge about the characteristics of the street addressing systems used in this region. These data were drawn from the historical transaction records of the Texas A&M GeoServices website (Texas A&M GeoServices 2013) which contained over 50 million individual address queries that have been sent to the production version of the system by members of the public (as described below). The specific records chosen for this research met one of two criteria: they had a ZIP code listed as being within one of those valid for Los Angeles City, or they had a city named 'Los Angeles'. Figure 3.2 shows the research area and the city boundary of Los Angeles City.

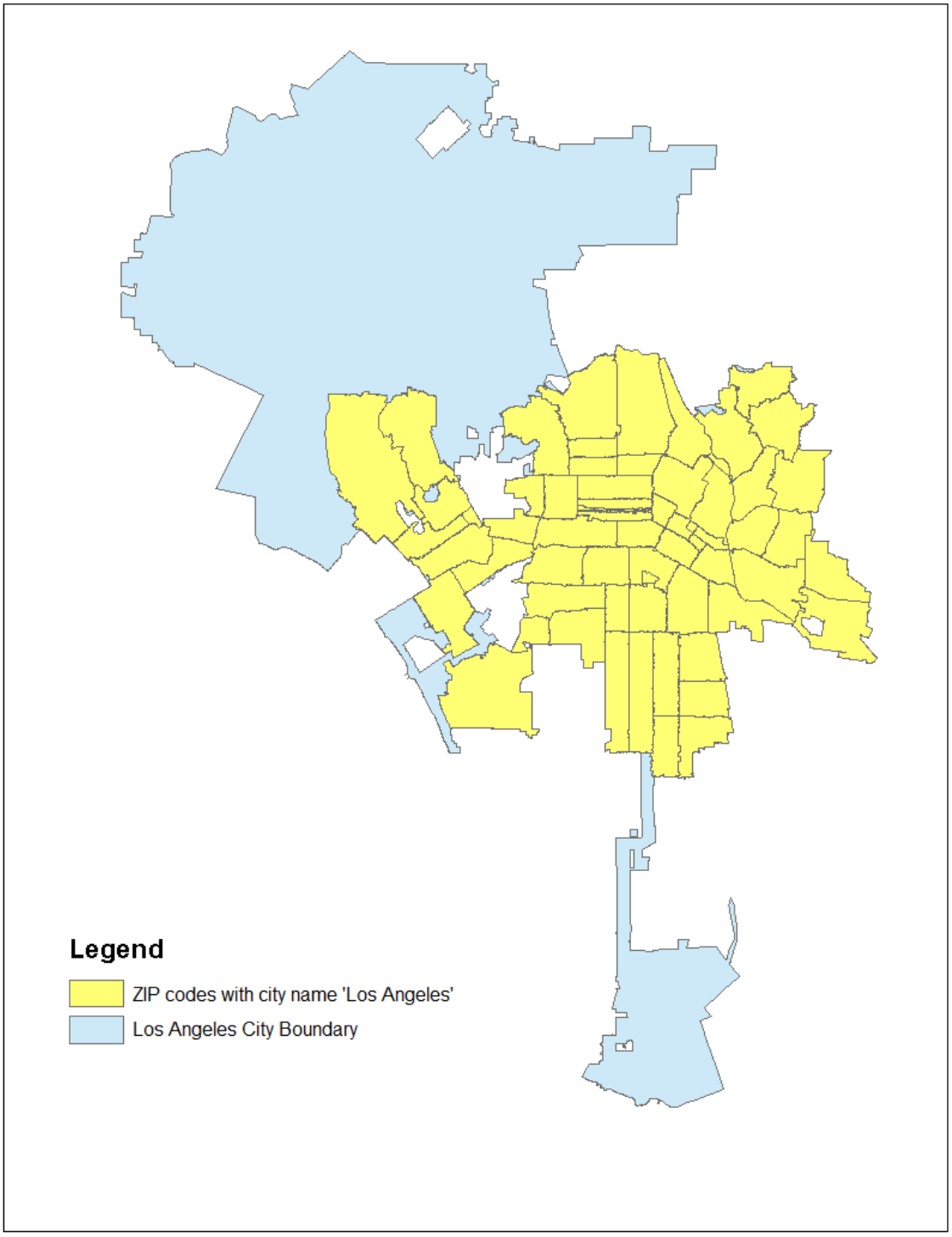


Figure 3.2 The research area and the Los Angeles City

3.2.2 Texas A&M GeoServices website

Figure 3.3 The website interface of Texas A&M GeoServices

The Texas A&M GeoServices website (Texas A&M GeoServices 2013) is an online platform that was developed by Dr. Goldberg and the Texas A&M GeoServices team. This website offers online services such as, geocoding, address parsing, and normalization, reverse geocoding, Census intersection and geocoding correction etc.

The Texas A&M GeoServices website processes hundreds of thousands of user input addresses every day. Among these input data, some data come from users who authorize Texas A&M GeoServices to use their data for research purposes. Based on these data, 19,273 records (input addresses) were selected as a test dataset for this research.

All test data were manual corrected with the help of the manual geocoding correction platform (Section 3.3). After this processing, all records had the following information associated with them: input address, corrected address, original geocoding result (probabilistic geocoding without region-based dynamic weighting method), and manually corrected longitude and latitude. While being processed, an addresses' error information was also collected. This error information was classified by the address field on which it occurred.

The Texas A&M GeoServices website was also used to represent a global-weighting probabilistic geocoding system as a comparison with the region-based dynamic weighting probabilistic geocoding system developed and evaluated as part of this research.

3.3 Manual geocoding correction

As part of the current study, a manual geocoding correction platform previously developed (Goldberg et al. 2008) was enhanced to gather additional information about the correctness of a geocoding result and geocoding input data.

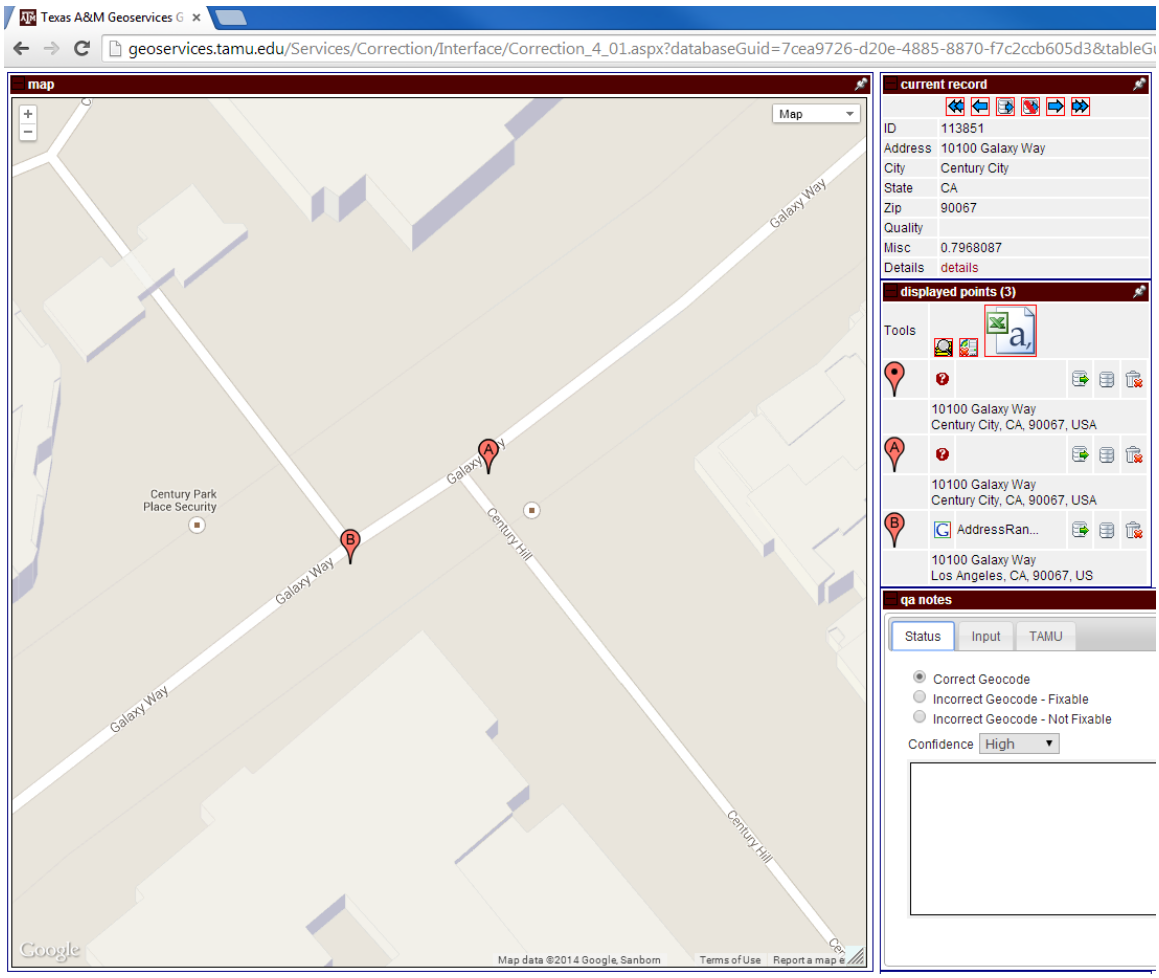


Figure 3.4 The interface of the manual geocoding correction platform

Figure 3.4 shows the interface of the manual geocoding correction platform. It was a webpage developed using the Google Maps API (application programming interface). It allowed users to manually correct the geocoding results in an effective and time efficient manner. For each input address, both the Texas A&M geocoding result point and the Google geocoding result point (or points in many cases) were displayed on a Google Maps interface. Input addresses and matched feature addresses from the multiple platforms were compared. Customized correction information was gathered using this approach. This

platform allowed users to drag and drop these point markers on the map to manually correct longitude and latitude coordinates.

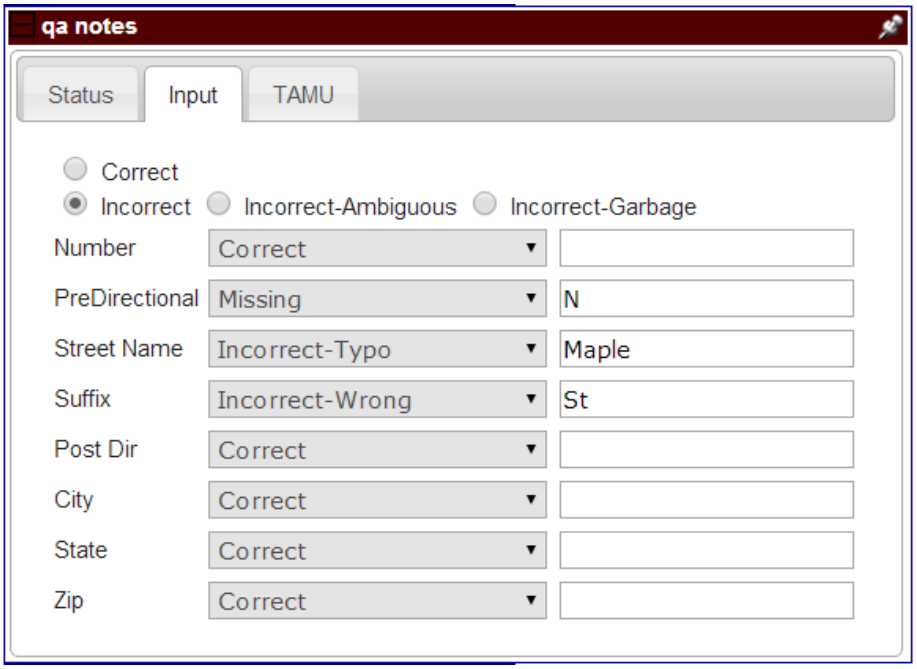


Figure 3.5 The geocoding correction notes

All records were compared with the matched feature in the reference database and Google geocoding results using the manual geocoding correction platform. New, corrected, addresses were generated during this process. For example, input address '10100 Galaxy Way, Century City, CA 90067' was compared with the computed geocode for '10100 Galaxy Way, Los Angeles, CA 90067' and Google geocoding result '10100 Galaxy Way, Los Angeles, CA 90067'. In this instance, the map interface indicated that all the address points in this ZIP code area (90067) were associated with the official city name 'Los Angeles'. 'Century City' is a colloquial term that is commonly used by people in this sub-region, but is not an official city name. Based on these facts, the address would be corrected

during manual processing to ‘10100 Galaxy Way, Los Angeles, CA 90067’ by the manual correction technicians. During the geocoding correction process, a roof-top longitude and latitude coordinate was also generated for each record. As Figure 3.5 indicates, error information at the per-attribute level was also collected for the original address, the version matched by Google, and the version matched by the Texas A&M GeoServices website (global-weighting).

3.4 Region-based dynamic weighting probabilistic geocoding system design

Figure 3.6 illustrates the workflow design for the region-based dynamic weighting probabilistic geocoding system.

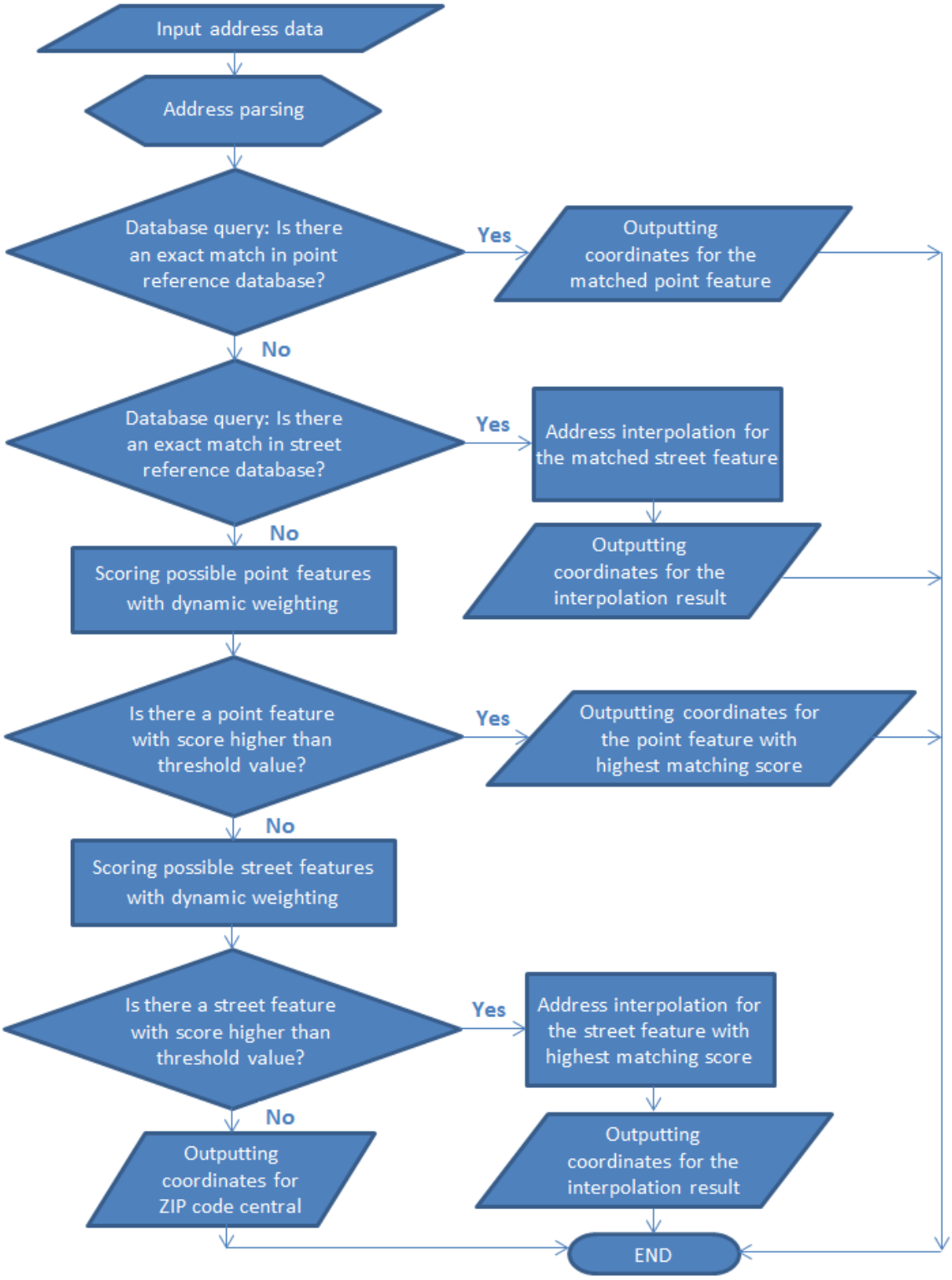


Figure 3.6 The region-based dynamic weighting probabilistic geocoding workflow

It is well known within the geocoding literature that even the best interpolation algorithms may still result in inaccurate or erroneous geocoding results. However, there is consensus that an exact match will always better than probabilistic matches. As is common in geocoding systems and geocoding research, the priority of reference datasets and the matching options were ranked as follow: 1) an exact match from the NAVTEQ Address Point Database; 2) and exact match from the NAVTEQ Street Segments Database; 3) a relaxed match (with region-based dynamic weighting) from the NAVTEQ Address Point Database; 4) a relaxed match (with region-based dynamic weighting) from the NAVTEQ Street Segments Database; and 5) a ZIP code centroid. Parsed input addresses were compared with reference database following this ordering.

In order to reach the objective of this study, including the construction of a geocoding system which implements region-based dynamic weighting method, following steps were accomplished.

3.4.1 Reference database preparation

NAVTEQ has been cited one of the best street address databases commercially available and has been widely used as reference dataset by many popular geocoding systems (Vieira et al. 2010). It provides coverage and accurate address data worldwide (Ludwig et al. 2011). Both NAVTEQ Address Point Database and NAVTEQ Address Point Database were used as reference datasets within the geocoding system developed and tested here.

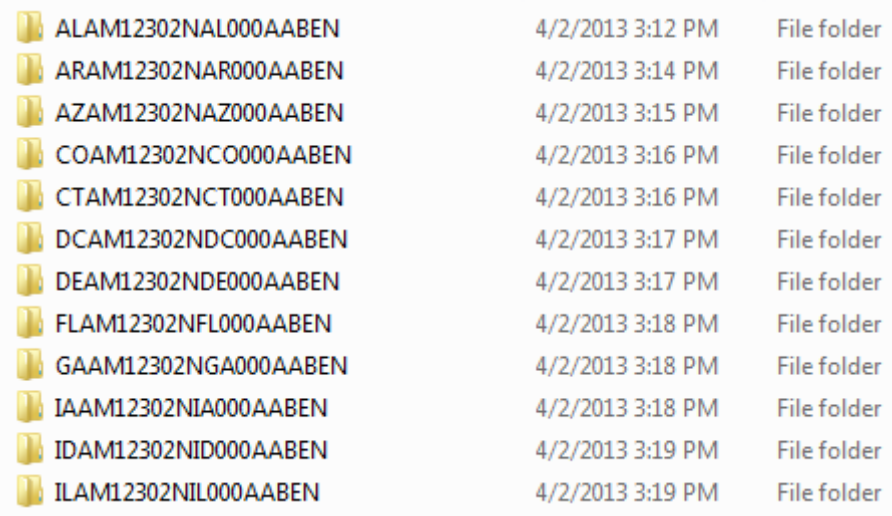


Figure 3.7 NAVTEQ Street Segments Database's folders for each state

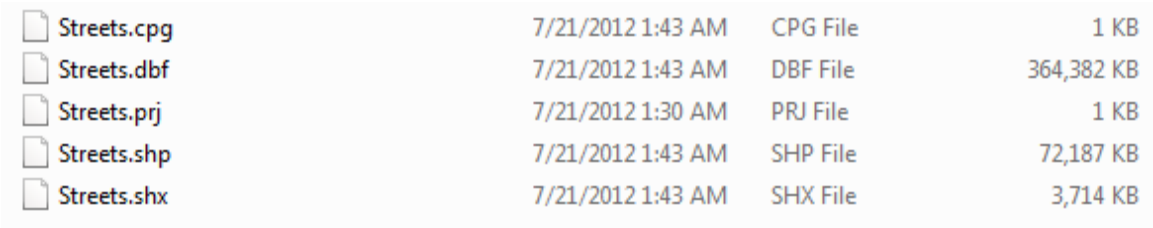


Figure 3.8 NAVTEQ Street Segments Database's shapefile for one state

As the Figure 3.7 and Figure 3.8 shows, the NAVTEQ Street Segments Database is delivered as one shapefile for each state. The shapefile format, an ESRI standard, is a widely used spatial vector data format designed by ESRI for GISystems (ESRI 1998). This format stores non-topological geometry data and attribute information for spatial features (i.e., points, lines, and polygons). Depending on the spatial size of the state and the density of the streets it contains, the size of the shapefile may vary from thousands to millions of records.

In order to be used as a reference dataset for the geocoding system, the NAVTEQ Street Segments Database and NAVTEQ Address Point Database was imported into a Database Management System (DBMS) to enable key functions like querying, updating, indexing, etc. In this research, Microsoft SQL Server (SQL Server) was used as the Relational Database Management System (RDBMS) that stored and managed these reference datasets.

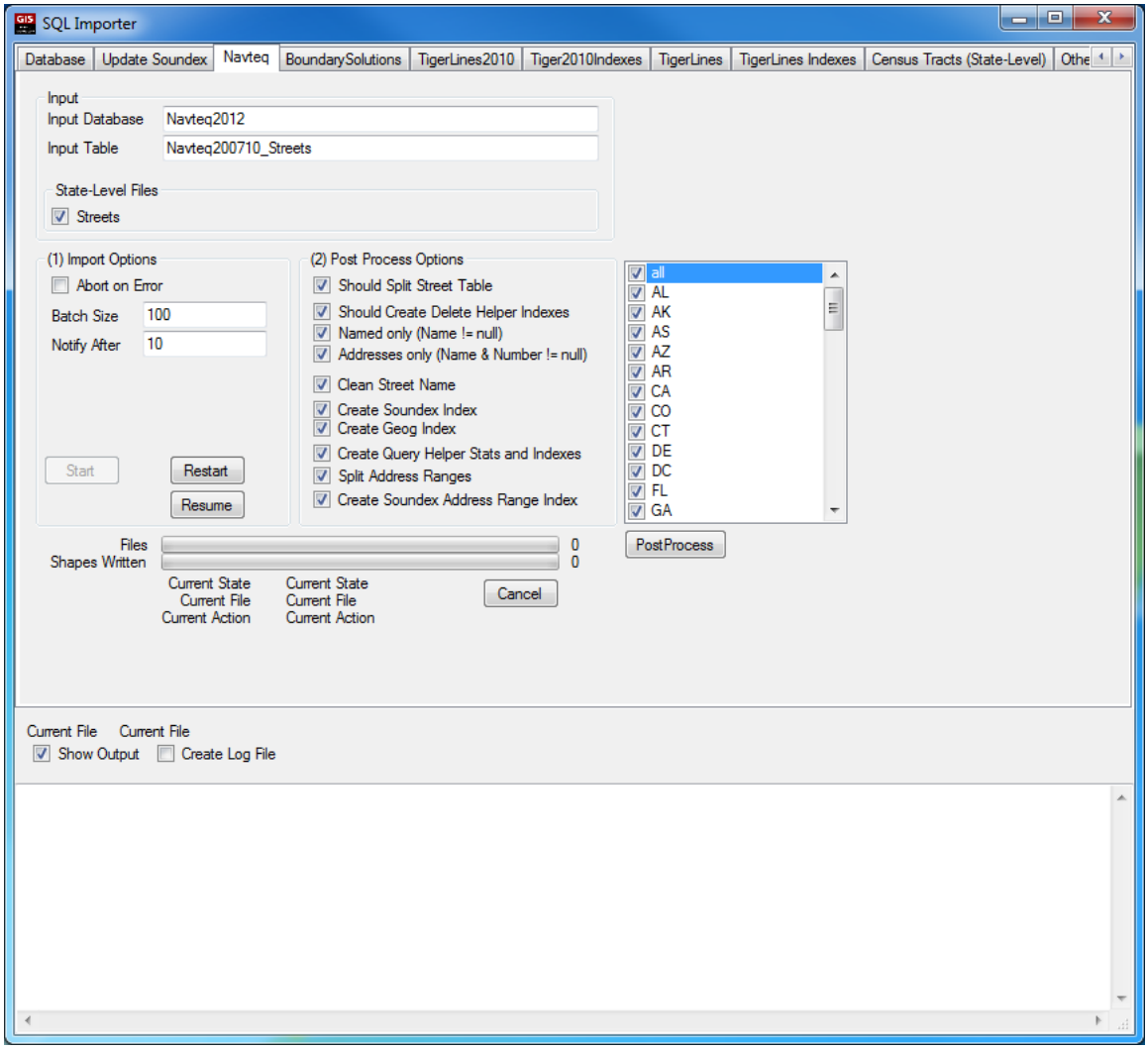


Figure 3.9 The NAVTEQ SQL Importer

Several pre-processing steps were run on the NAVTEQ data in order to build a fully functional and effective reference dataset for the geocoding system used here. In this study, the NAVTEQ SQL Importer (Figure 3.9) was developed as a tool for data importing and pre-processing. It could read each of the shapefiles for each state separately from its sub-folder and import them into SQL Server. As the data were being imported, several processing steps were executed upon them to ensure proper formatting as well as automatically prepare the geocoding system for the approaches outlined below.

As mentioned in Section 2.1.3, address interpolation calculates the left-hand street and the right-hand street separately based on the parity of the address number. Therefore, separating the addresses by left-hand and right-hand (even/odd) in each state significantly improved the query efficiency by reducing the searching range and parity determination time requirements.

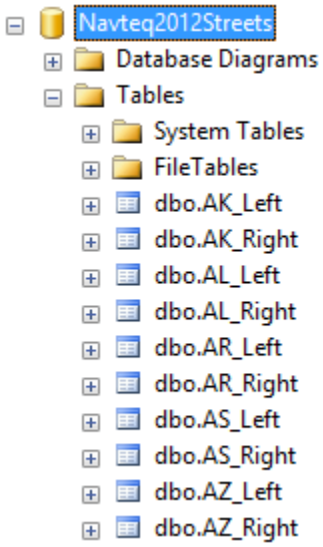


Figure 3.10 The results of addresses left-right separation

Figure 3.10 shows the results of addresses left-right separation. All addresses in the NAVTEQ Street Segments Database for the US were inserted into independent tables by state, and by left-right designation.

The original NAVTEQ data did not include city name and state name associations with each street record. However, each street line record had both left-hand postcode (ZIP code) and right-hand postcode data. Based on the ZIP code database provided by the USPS (the United States Postal Service), all the city and state names were added into the database during the data importing process.

	name	nameSoundex	cityNameLeft	cityNameLeft_Soundex
13028	KOGRU	K260	Eagle River	E246
13029	KUCERA	K260	Big Lake	B242
13030	KIZER	K260	Clam Gulch	C452
13031	KUGRUPAGA	K261	Brevig Miss...	B612
13032	KODY	K300	Fairbanks	F615
13033	KATYA	K300	Fairbanks	F615
13034	KATHY	K300	Anchorage	A526
13035	KODY	K300	Anchorage	A526
13036	KATHY	K300	Homer	H560
13037	KATIE	K300	Palmer	P456
13038	KATIE	K300	Palmer	P456

Figure 3.11 Samples of Soundex code fields and their original data fields

As discussed in Section 2.1.5, the Soundex encoding technology is commonly used by geocoding systems to solve misspelling errors that exist in input data. The NAVTEQ SQL Importer pre-computed the Soundex code of street name and city name fields for each record as they were being imported. Pre-computed Soundex codes were stored in extra fields in the reference database along with the original records. These pre-computed fields were used to increase the real-time processing time efficiency of the system by reducing

the time needed to run the encoding algorithms in real-time. When the Soundex codes were pre-computed and stored in the reference database, only the Soundex of the input address needed to be computed for comparison against the reference data.

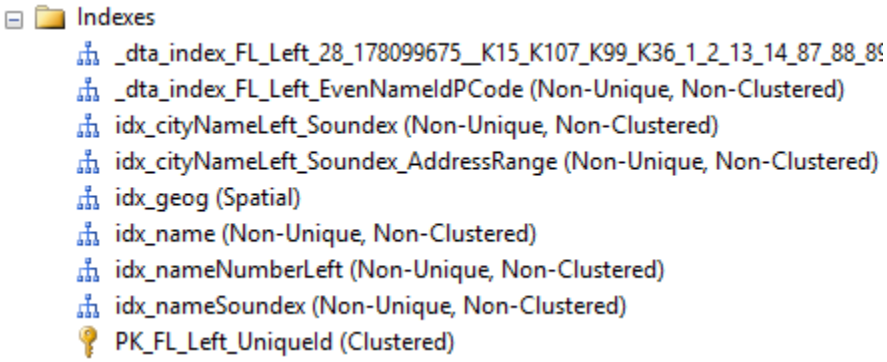


Figure 3.12 The pre-generated indexes examples for each table

As Figure 3.12 demonstrates, both spatial and non-spatial indexes that were related to the most frequently used queries were generated by the NAVTEQ SQL Importer during the importing process.

After the conclusion of the importing process and the reference data preparation steps mentioned above, the NAVTEQ Street Segments Database was converted into a ready-for-use SQL Server reference database. Importing the NAVTEQ Street Segments Database followed a similar process to that just described.

3.4.2 Dynamic weighting regions generation

As noted, a globally-defined attribute weighting scheme fails to capture the local street variations, distribution, conventions, and characteristics of a region. The purpose of the current research was to develop an approach that would utilize this locally-based knowledge to improve the accuracy of geocoding systems. To accomplish this, two tasks

needed to be completed. First, the concept of a local street ‘signature’ needed to be developed which was capable of representing the characteristics of the streets within a local region. Second, an automated approach to determining a ‘local region’ needed to be developed such that the ‘signatures’ could be derived (or more aptly ‘grown’) based on the characteristics of streets in a region.

3.4.2.1 Street signature for ZIP code areas

To accomplish the first of these tasks, the spatial unit of a ZIP code (or more precisely, an approximation of area based on the delivery route of the USPS [Grubesic et al. 2006]) was used as a first order descriptor of a ‘local region’ within which to discern the ‘character’ of streets. Street signatures within ZIP code areas were designed as a set of attributes which identified the character of streets in distinct ZIP code area. Street signatures were generated based on the values and density of address component types and values in each ZIP code area. In this study, percentage of pre-directional, percentage of post-directional, percentage of numerical street name (e.g., First St and 17th Ave) and percentage of Spanish street name (e.g., San Pedro St and La Cienega Blvd) were used as main parameters of a street signature. The NAVTEQ Street Segments Database was used to collecting these attributes and generate signatures.

Table 3.1 Definitions of each street signature component

Attribute	Definition (Calculation)
Percentage of pre-directional	Number of streets which have pre-directionals / Number of total unique streets in one ZIP code area
Percentage of post-directional	Number of streets which have post-directionals / Number of total unique streets in one ZIP code area
Percentage of numerical street name	Number of streets which are numerical street names / Number of total unique streets in one ZIP code area
Percentage of Spanish street name	Number of streets which are Spanish street names / Number of total unique streets in one ZIP code area

In this approach, one 'street' was defined by a unique combination of pre-directional, street name, street suffix, and post directional. Therefore, each attribute of the street signature was defined as listed in Table 3.1. All calculations and counting were achieved by SQL (Structured Query Language) within SQL Server and indexed per unique ZIP code. Figure 3.13 demonstrates the examples of street signature sets for ZIP code area.

	ZIPcode	Lat	lon	PerPre	PerPost	PerNum	PerSpan
1	90001	33.97331	-118.243612	0.5063291139	0	0.4177215189	0
2	90002	33.948432	-118.245196	0.4177215189	0	0.329113924	0
3	90003	33.962989	-118.272517	0.9384615384	0	0.7846153846	0.0153846153
4	90004	34.076333	-118.307713	0.7899159663	0	0.0168067226	0.0420168067
5	90005	34.055509	-118.30922	0.7049180327	0.0163934426	0.0655737704	0.0163934426
6	90006	34.045817	-118.290338	0.6440677966	0	0.1864406779	0.033898305
7	90007	34.026525	-118.282408	0.6363636363	0	0.2987012987	0.0129870129
8	90008	34.006715	-118.339229	0.1567164179	0	0.119402985	0.2462686567
9	90010	34.063478	-118.314966	0	0	0	0
10	90011	34.006146	-118.257237	0.5888888888	0	0.5	0.0333333333
11	90012	34.063679	-118.238854	0.4153846153	0	0.0307692307	0.0307692307

Figure 3.13 The examples of street signature set for ZIP code area

3.4.2.2 Dynamic weighting regions consolidation

The second step in the approach was to automatically grow regions which had similar street signatures into larger regions which shared the same street ‘character’. These weighting regions were generated by combining a set of adjacent ZIP code areas that had similar street signatures.

This region consolidation approach was accomplished by grouping adjacent zip codes with similar street signatures. Street signature similarity was determined by comparing the classified values for each component of the street signature. ZIP codes that contained the same classification results of each of the signature fields, and were spatially adjacent, were consolidated into a single region. Class membership for street signatures were treated differently based on the type of attribute under consideration (Table 3-1), but all classes were derived using Jenks optimization method, also known as Jenks natural breaks classification method - a data clustering method designed to determine the best arrangement of values into different classes.

3.4.2.3 Region statistics database table

To speed up the processing time of the standard weight for each field, the statistical information for all the regions were pre-calculated and stored as a reference table in SQL Server.

RegionID	TotalNum	TotalNam	TotalType	TotalPre	TotalPost
1	3576	268	14	5	3
2	2625	156	10	3	3
3	7310	419	12	5	1
4	269	8	7	3	3
5	3182	190	15	3	2
6	4578	205	9	3	1
7	2364	89	8	3	1
8	1670	128	7	3	1
9	2923	116	6	3	1
10	5575	326	13	4	2

Figure 3.14 The examples of statistics for region

As Figure 3.14 shows, the total number of unique address numbers, total number of unique street names, total number of unique street types, total number of unique pre-directionals, and the total number of unique post-directional were pre-calculated for each region.

3.4.3 Weighting calculation

The real-time generation of statistics for all attributes in every field of the address components for all reference features would cause large amounts of redundant work and intolerable processing time. As such, the weighting scores for address fields in each region were pre-computed by assuming the input value as a true match with the value of one record in the reference dataset (m=1) following the process outlined in section 2.1.4:

$$\text{Weight} = \ln(m/u) \tag{3.7}$$

$$\text{u} = 1/Total\ of\ Unique\ values\ in\ this\ field \tag{3.8}$$

Assuming the input street name is a true match with the street name of one record in the reference database in this region. (m = 1)

$$Weight_{Standard} = \ln(1/u) \tag{3.9}$$

This approach is best understood through the use of an example. Assume that in one region, there are 250 unique street names. Further assume the input street name is a true match with the street name of one record in the reference database in this region. This results in the probability (u) that one existing input street name to be a true match with a random street name in this region is 0.004 (Equation 3.8). In this case, the standard weighting for the street name field would be 5.52 (Equation 3.9).

In each case of the matching attempt, real weighting scores were calculated based on the possibility (m) of one input field being a true match with the trial record's field and the standard weight.

$$\text{Weight} = \ln(m/u) = \ln\left(\frac{1}{u}\right) + \ln(m) = Weight_{Standard} - \ln(\frac{1}{m}) \tag{3.10}$$

The process of searching in one region, based on the ZIP code of input address, assumed there was one record matching with the input address. Therefore, the possibility (m) of one input field being a true match with the trial record's field could not be lower than the probability (u) of two fields agreeing at random. So the weight was always positive.

3.4.4 Feature scoring

Figure 3.15 illustrates the workflow of feature scoring. This algorithm queried the database twice. The first query was to obtain the region ID of the address based on ZIP code. The second query was used to obtain the statistical data for the requested region. The weighting score was calculated based on the algorithm discussed previously (Section 3.4.3).

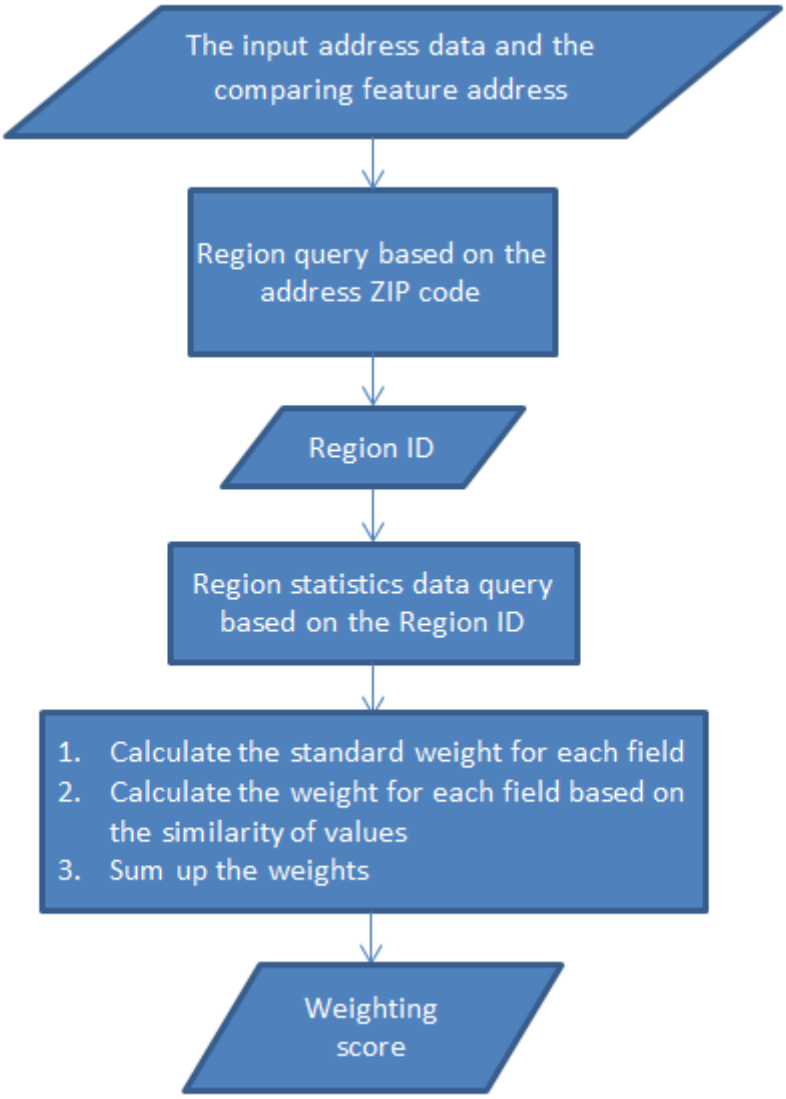


Figure 3.15 The workflow of feature scoring

4. RESULTS AND ANALYSIS

4.1 Results of manual geocoding correction

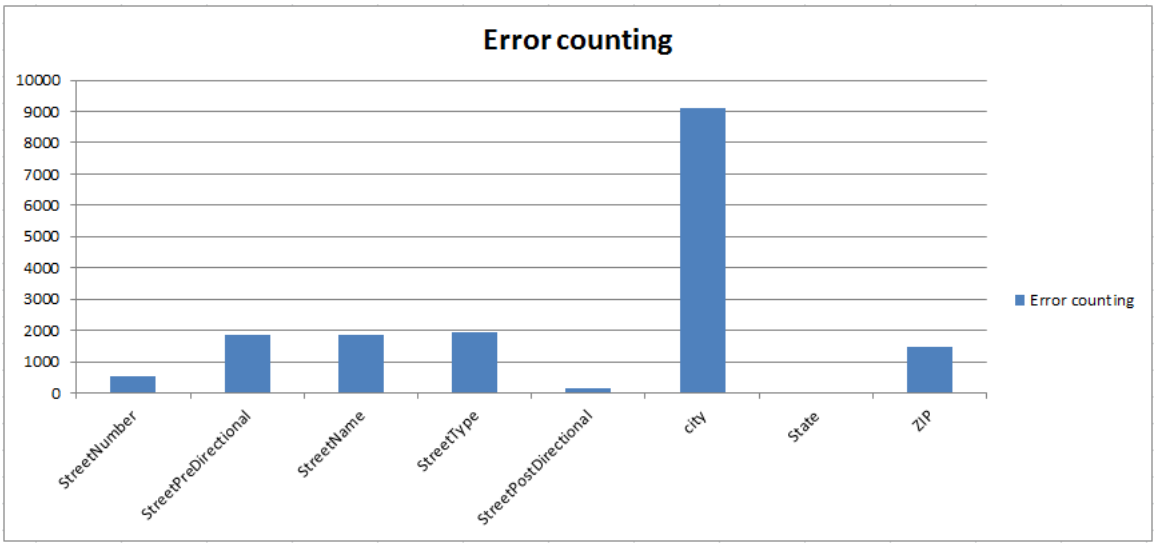


Figure 4.1 The error counting histogram

Figure 4.1 shows statistics of the resulting error information collected during the manually geocoding correction process. As expected, the city name field contained the largest amount of error. Particularly in Los Angeles City, many input addresses used a colloquial sub-city name (community name) instead of 'Los Angeles'. Missing city names were also an issue for input addresses.

4.2 Result of signature and region generation

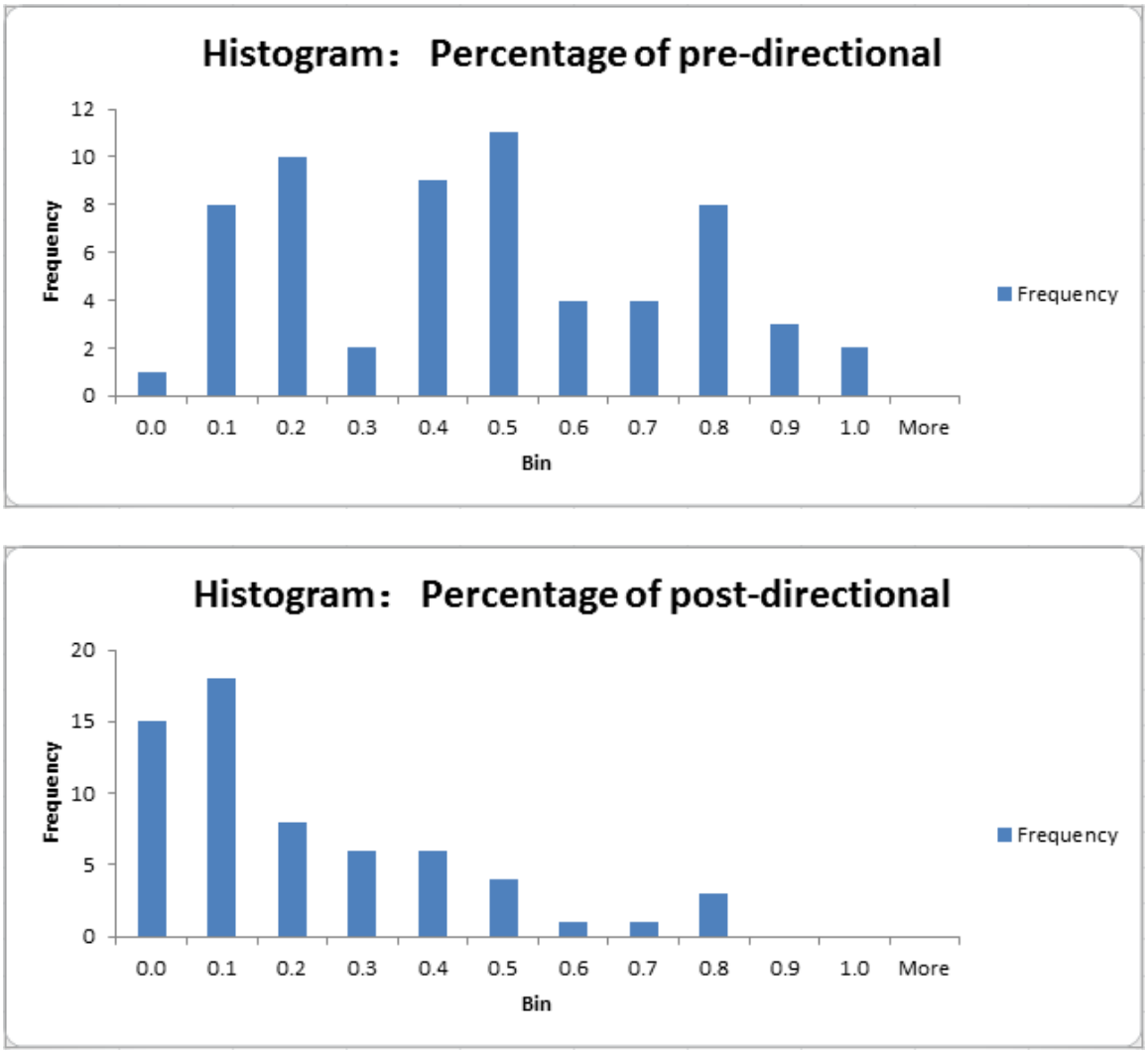
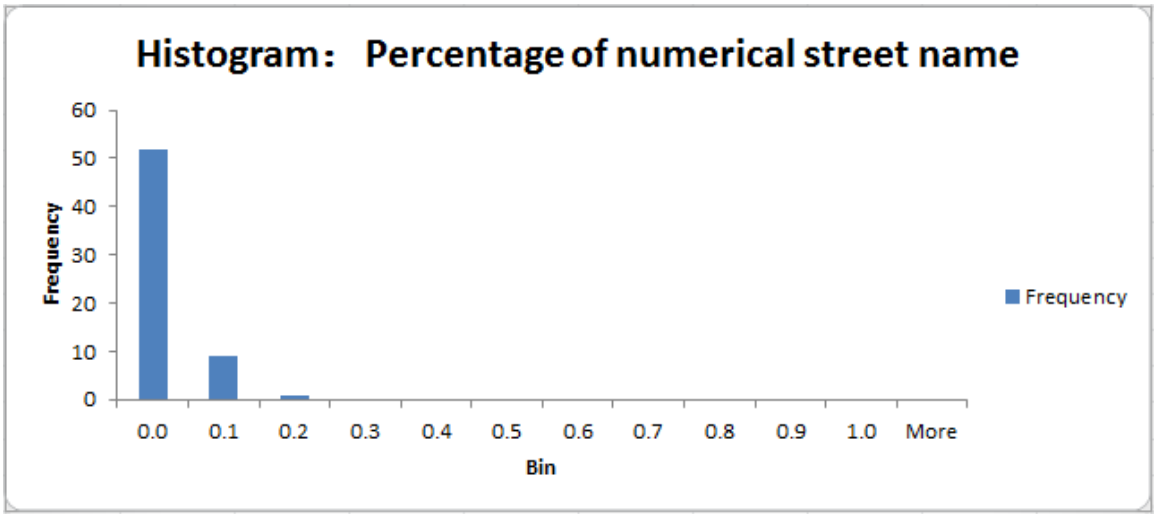


Figure 4.2 Histograms for street signature parameters



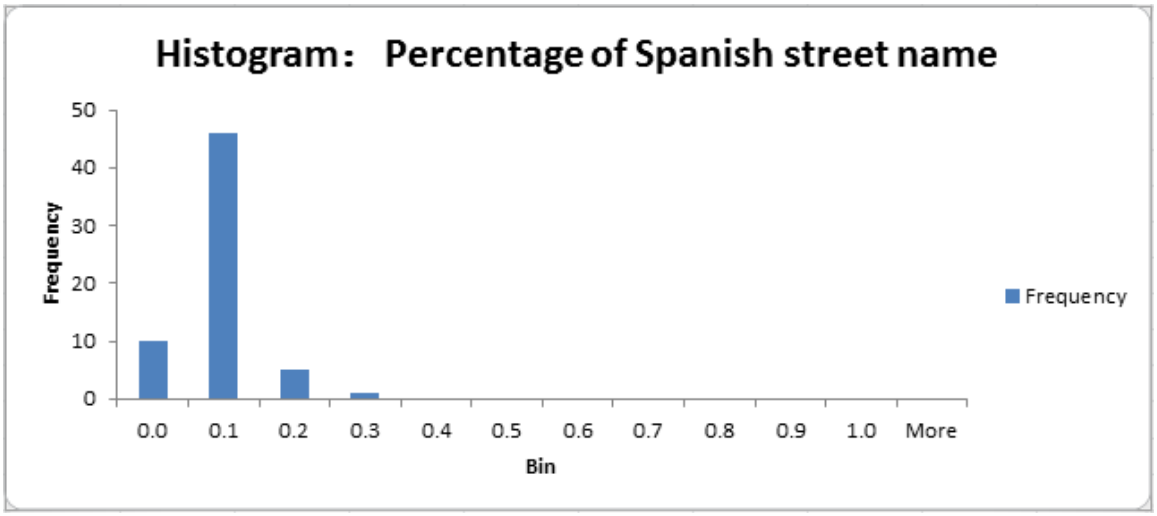


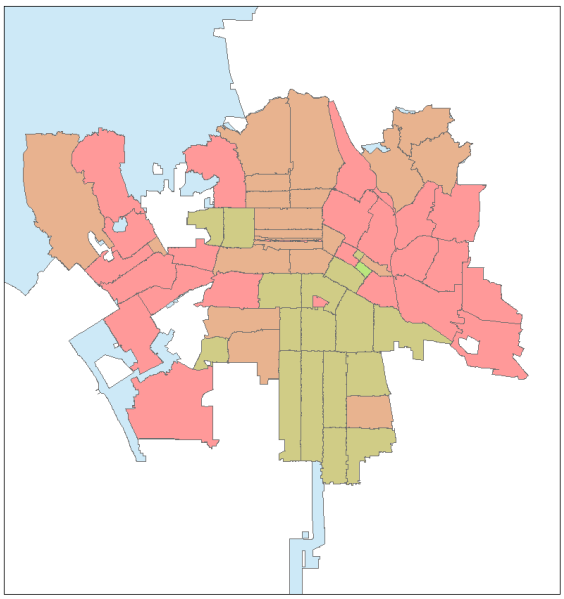
Figure 4.3 Histograms for street signature parameters

Figure 4.2 and Figure 4.3 shows four histograms for street signature parameters. As these histograms show, the parameters of the street signature are strongly inhomogeneous. Therefore, each parameter of the street signature was classified (2 classes, 3 classes, and 4 classes, respectively) by Jenks optimization method.

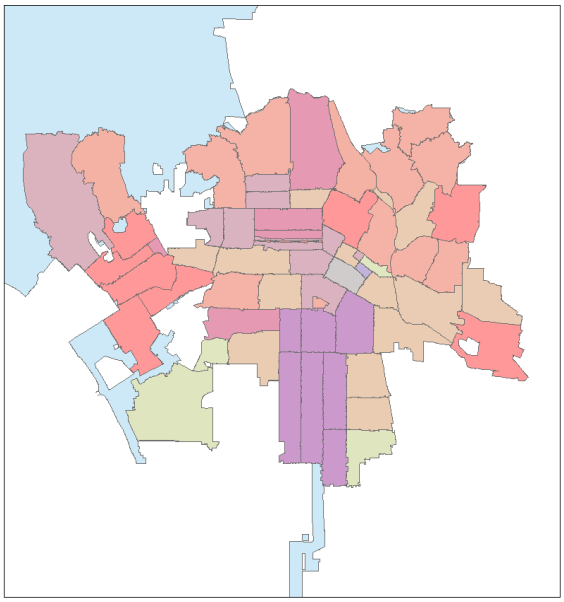
In order to evaluate the correctness of the region segmentation algorithm based on street signatures, the resulting regions were compared against neighborhood boundaries in

Los Angeles City. This evaluation technique leveraged the fact that communities in Los Angeles City, such the South Los Angeles, have street naming authority, and attempt to maintain a consistent street character for the regions they control. If the proposed method generated similar results to those which make up the borders of the neighborhoods Los Angeles City, it would mean that the automated approach to agglomerate areas which have similar street characteristics was capable of accurately mimicking the true distribution of heterogeneous street naming conventions. To evaluate this approach, a set of comparison maps were generated which overlaid all street signature parameters in different levels classification.

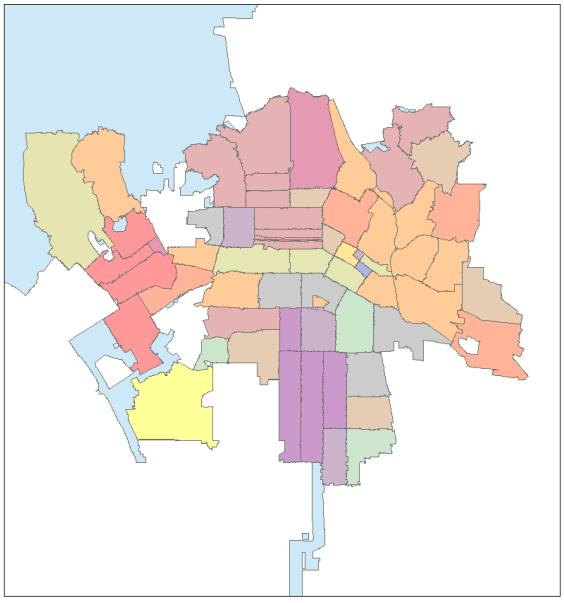
The accuracy of the region growing approach using three different classifications was evaluated by comparing to a known map of LA neighborhoods (Los Angeles Almanac 2004). Figure 4.4 (a, b, and c) shows the mapping results (2, 3, and 4 classes) of regions in Los Angeles City based on the similarity of street signatures in adjacent ZIP code areas. The Los Angeles Communities map from Los Angeles Almanac (2004) used as the base area comparison is displayed in Figure 4.4 (d). The regions generated by street signatures that are classified into 3 classes resulted in the highest degree of similarity, both in terms of overall shape and representative size, as compared to the Los Angeles Communities map. Based on the semi-official version above, regions that were generated by the 3-classes classified street signatures were deemed the closest match and selected as the dynamic weighting regions for the research area.



(a) 2 classes



(b) 3 classes



(c) 4 classes

(b) Los Angeles Communities map in research area

(Los Angeles Almanac 2004)

Figure 4.4 The region maps (2, 3, and 4 classes) and Los Angeles Communities map

4.3 Experimental result

Of the total records manually corrected, 5,379 records were geocoded by both the global-weighting geocoding system and the region-based dynamic weighting probabilistic geocoding system. The improvement (*I*) was calculated for each record based on the difference between errors of both the old and the new algorithms. Based on the definition of improvement index above (Equation 3.6), a positive value represented an improved result and a negative value represented a degraded result.

4.3.1 Results in spatial accuracy

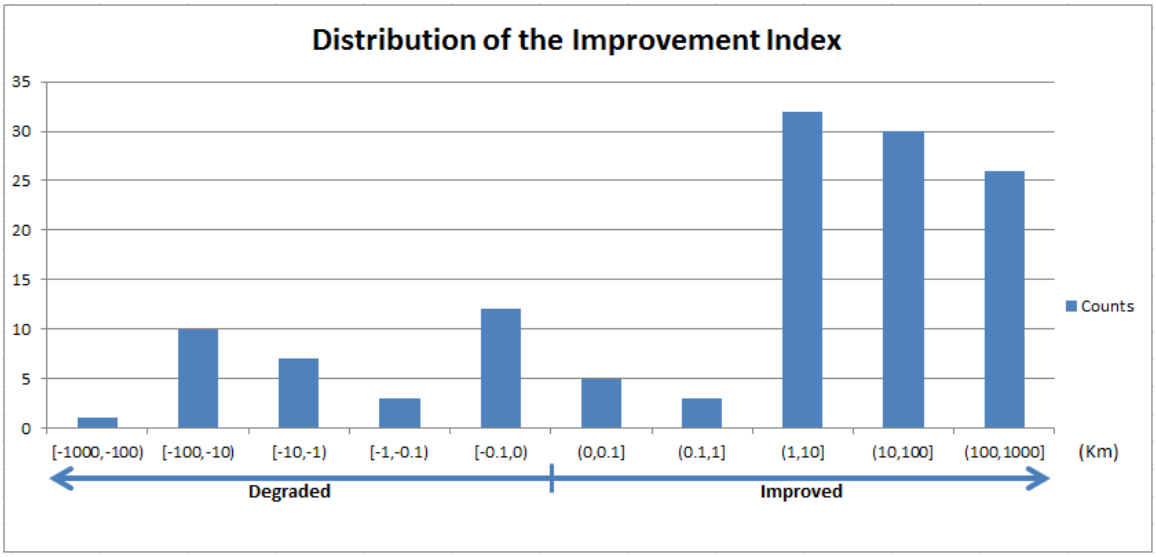


Figure 4.5 The distribution of improvements index

Figure 4.5 shows the distribution of improvement index. 96 records (1.78%) were improved while 33 (0.61%) records degraded. Most of the improvements (88, 92%) were greater than 1 kilometer, meaning that the proposed region-based dynamic weighting method improved spatial accuracy by 1 kilometer over a geocoding approach which used

a global-weighting of address attributes (non-dynamic). By summing all the improvement indexes (both positive and negative index), the new system improved the geocoding results by 7,562.36 kilometers in total.

87 records failed to match in the global-weighting geocoding system because they were given scores that were lower than the threshold value (88 match score) and were thus matched to a lower level of geography (ZIP code centroid) instead of matching at an address point or street segment.

The weighting changes resulting from the region-based dynamic weighting versus the global-weighting probabilistic weighting show in Figure 4.6.

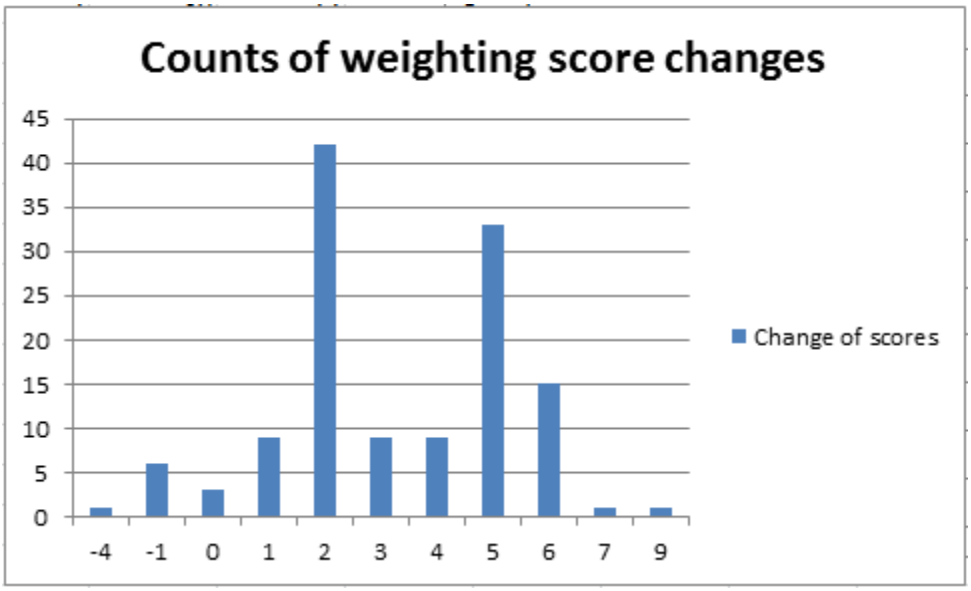


Figure 4.6 The counts of weighting score changes

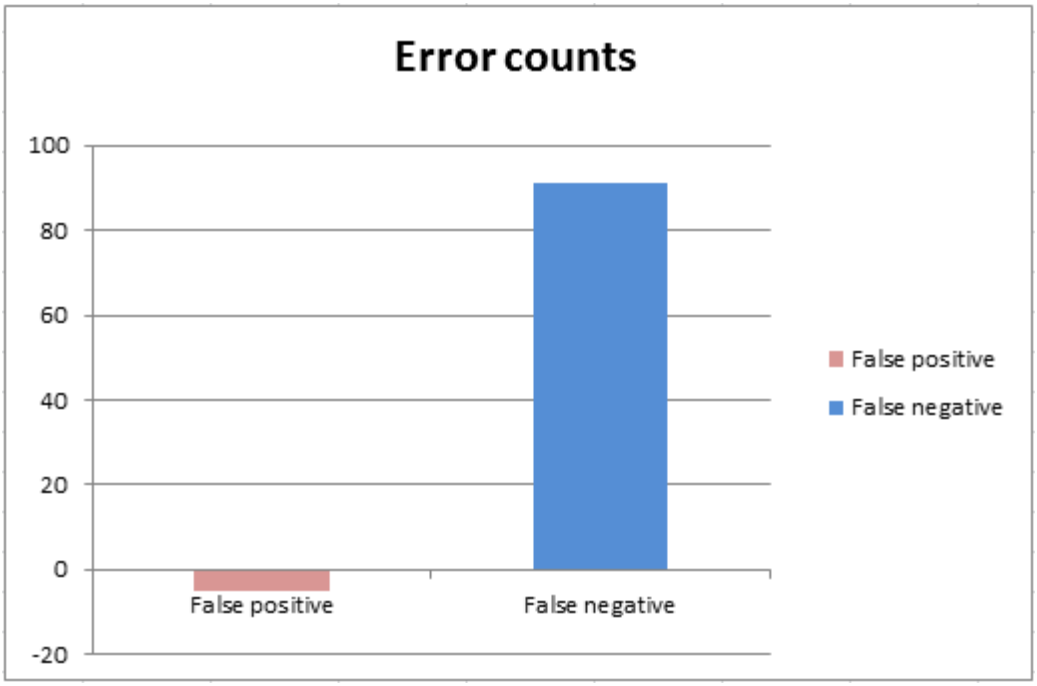


Figure 4.7 Error counts of the global-weighting probabilistic geocoding system

Figure 4.7 shows the error counts of the global-weighting probabilistic geocoding system. 96 records (1.78%) improved using the region-based dynamic weighting probabilistic geocoding system. 5 of those records failed to generate correct geocoding results in the global-weighting probabilistic geocoding system because of the false positive errors. Their original weighting scores computed with the global-weighting method were higher than the scores with the proposed method. 91 of records (1.69%) failed to generate correct geocoding results in the global-weighting probabilistic geocoding system because of false negative errors. Their original weighting scores computed with the global-weighting method were lower than the scores with the proposed method.

For example, '10375 WILSHIRE BLDV Los Angeles, CA 90024' was geocoded to ZIP code centroid in the global-weighting probabilistic geocoding system because the

misspelled suffix name ‘bldv’ was recognized as part of the street name ‘Wilshare BLDV’ instead of suffix filed ‘Blvd’. This resulted in a match score which was lower than the threshold value (88 match score), and thus an ultimate match at the ZIP code level rather than the address point or street range. In the new system, this input address was computed to have a match score of 88.39 when compared to the record ‘10375 Wilshire Blvd Los Angeles, CA 90024’ in NAVTEQ Street Segments Database, resulting in an improved geocode output (an address point rather than a ZIP code centroid).

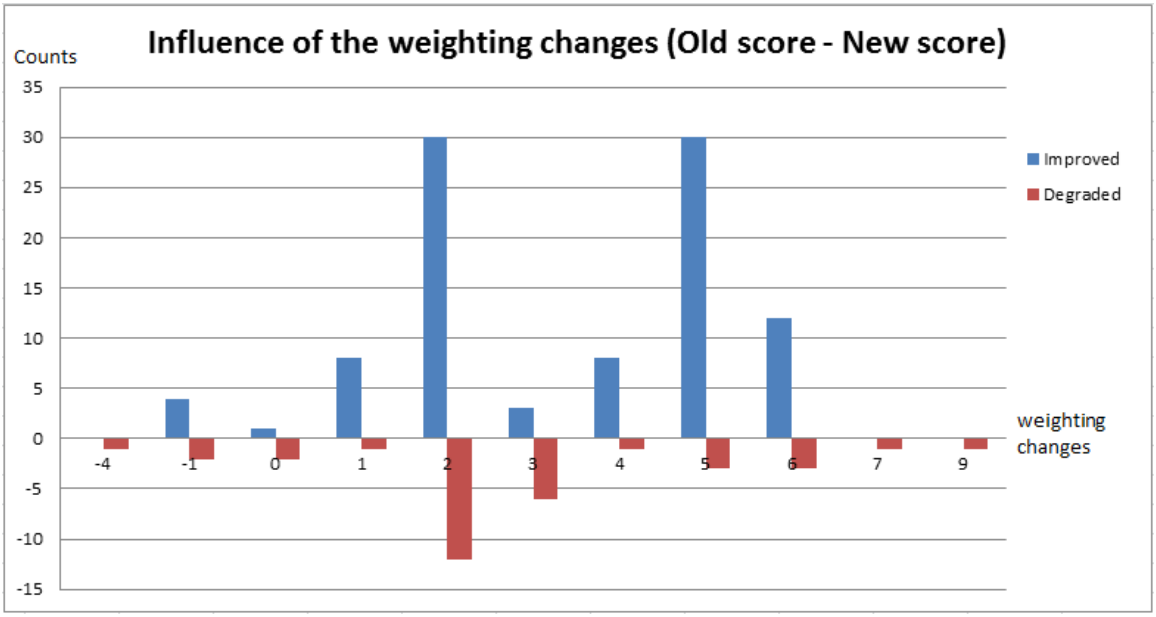


Figure 4.8 The influence of the weighting changes

Figure 4.8 shows the influence of the weighting changes on those records that were observed to have differences in match scores between the global-weighting and the region-based dynamic weighting approaches. When compared with the global-weighting probabilistic geocoding system, results indicated that the majority of the weighting scores

for region-based dynamic weighting probabilistic geocoding system were higher, indicating that the system improved the false negative rate of the geocoding test dataset. The range of weighting changes (1 to 6) appears to have had the best influence to geocoding accuracy.

4.3.2 Results in time efficiency

The batch geocoding process using the global-weighting probabilistic geocoding system over 19,273 records completed 1 hour and 47 minutes. In contrast, the same input dataset completed in 2 hour and 23 minutes using the region-based dynamic weighting approach. The main increase in processing time was due to the need to issue two extra requests per input query for each record during the region-based dynamic weighting process. However, for certain records, the region-based dynamic weighting probabilistic geocoding system resulted in faster processing times when compared to the global-weighting system. This was due to increased match rates resulting from matches being found in a higher priority dataset, meaning that the geocoding system did not have to search as many datasets since a match was found using the new approach when previously the geocoding system had to search through a larger number of datasets to find a solution. For example if the matching algorithm found a match in the NAVTEQ Address Point Database, the system would not need to subsequently search through the NAVTEQ Street Segments Database and ZIP code centroid database.

The increased processing time of region-based dynamic weighting probabilistic geocoding system was deemed acceptable for most geocoding tasks. Because this study only tested the proposed method in Los Angeles City, the algorithm needed to execute an additional query to ensure that the input address fell within in the test area of Los Angeles

City every time an address was to be geocoded. When applying this method for the whole country in future work, this restriction would be removed and the number of requests for obtaining region information in real-time would likely reduce to one. Therefore, the processing time would be faster than observed in the present implementation.

4.4 Discussion

In this experiment, manually corrected geocoding results were used as 'gold standard' coordinates to evaluate a new geocoding technique. These manually corrected geocoding results produced using a combination of Google geocoding results, a global-weighting probabilistic geocoding system's results, and human estimated results. Each of the above datasets and processes had its own inherent limitations and biases which are reflected in its output data. In some cases, the manually corrected geocoding results did not represent perfect 'gold standard' points. In some instances, this may have resulted in a biased evaluation of the region-based dynamic weighting probabilistic geocoding system. The following list of specific cases exemplifies specific problems that were observed. These instances motivate future work to refine the method presented here even further.

Case 1: Directional/ZIP code confusion: For the input address '1357 W Vernon Ave LA,CA 90011', both Google Maps and the global-weighting probabilistic geocoding system produced points that referred to the address '1357 E Vernon Ave LA,CA 90011'. However, the region-based dynamic weighting probabilistic geocoding system output a point that refers to the address '1357 W Vernon Ave LA, CA 90037' (which upon review was determined to be a better match). In such cases, it was extremely difficult to tell where the original input should have been located. Either choice of the output point would

generate a different evaluation result when compared to that generated by the proposed system.

Case 2: Pre/Post-directional/Name/City Confusion: ‘10559 eastern Ave west Hollywood CA 90064’ represented an ambiguous input addresses. The city name of the input address ‘West Hollywood’ was actually a sub-area name of Los Angeles City, and ‘eastern Ave’ could not be found in the ZIP code listed with the address. In this case, both the Google geocoder and the global-weighting probabilistic geocoding system could not find a match for this address. Therefore, the gold standard coordinates were set as the geometry centroid of the ZIP code area (a degradation of geocode accuracy since this falls below street level). In contrast, the region-based dynamic weighting probabilistic geocoding system matched this address to ‘10559 Esther Ave Los Angeles, CA 90064’ since ‘Eastern’ was recognized as a typo of ‘Esther’ and resulted in a match score of 90.35 which was above the minimum match threshold. Manual inspection revealed that this address match should be an acceptable match for the input address given the distribution of street names in the ZIP code and Los Angeles City. However, if an analysis of the results of the system only compared the manually corrected gold standard point, the region-based dynamic weighting probabilistic geocoding system would be judged as producing a degradation in quality for this record.

5. CONCLUSION AND FUTURE WORK

This thesis has designed, developed, and tested a new method for using region-based dynamic weighting probabilistic geocoding in order to increase the level of spatial accuracy and decrease the level of spatial uncertainty in geocoded data. Specifically, it has accomplished the following tasks. **Chapter 1 and 2** described the spatial errors present in geocoded data, and the effects on data and analyses that may result from ignoring the local differences between addressing and street naming conventions between regions as well as the need for accurate geocoding result in numerous research areas. **Chapter 3** described the experimental design workflow as well as preparation steps, like reference dataset importing, pre-processing, and region information pre-calculating, that were necessary in order to implement and evaluate the region-based dynamic weighting algorithm. **Chapter 4** evaluated the results of this algorithm in terms of spatial accuracy, match score change, and time efficiency of region-based dynamic weighting in comparison to results generated through the global-weighting probabilistic geocoding system.

The results of this study show that taking the place-specific naming conventions of different regions into account, a region-based dynamic weighting probabilistic geocoding system be constructed which improves the spatial accuracy of geocoding results. Based on an evaluation of 5,379 manually inspected and corrected records, the proposed method improved the spatial accuracy of geocoding results by 7,562.36 kilometer in total over the full research study area. Most geocoding errors generated by the global-weighting probabilistic geocoding method in this study area were due to false negative errors.

Although the time efficiency of the new method degraded, it was deemed acceptable for most geocoding processing tasks.

In future work, this method will be applied across the entire US, and its results will be evaluated using testing data from across the country. The problem of ambiguous ‘gold standard points’ test data will be investigated through additional data generation methods such as error simulations for input address data. These test data will be generated by selecting point address records from reference databases and modifying the text address fields by randomly adding common errors resulting in test data that have text addresses with known errors associated with correct spatial points.

REFERENCES

Amram, O., Abernethy, R., Brauer, M., Davies, H., & Allen, R. W. (2011). Proximity of public elementary schools to major roads in Canadian urban areas. *International Journal of Health Geographics*, *10*(68), 1-11.

Andresen, M. A. (2006). A spatial analysis of crime in Vancouver, British Columbia: a synthesis of social disorganization and routine activity theory.*The Canadian Geographer/Le Géographe canadien*, *50*(4), 487-502.

Bakshi, R., Knoblock, C. A., & Thakkar, S. (2004, November). Exploiting online sources to accurately geocode addresses. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems* (pp. 194-203). ACM.

Baker, J., Alcantara, A., Ruan, X., & Watkins, K. (2012). The impact of incomplete geocoding on small area population estimates. *Journal of Population Research*, *29*(1), 91-112.

Balmes, J. R., Earnest, G., Katz, P. P., Yelin, E. H., Eisner, M. D., Chen, H., ... & Blanc, P. D. (2009). Exposure to traffic: Lung function and health status in adults with asthma. *Journal of Allergy and Clinical Immunology*,*123*(3), 626-631.

Bell, B. S., Hoskins, R. E., Pickle, L. W., & Wartenberg, D. (2006). Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. *International Journal of Health Geographics*, *5*(1), 49.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology*, *14*(4), 408-412.

Boscoe, F. P. (2008). The science and art of geocoding: Tips for improving match rates and handling unmatched cases in analysis. *Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research and Practice*, 95-110.

Burrough, P. A., McDonnell, R., Burrough, P. A., & McDonnell, R. (1998).*Principles of geographical information systems* (Vol. 333). Oxford: Oxford university press.

Ceccato, V., & Haining, R. (2004). Crime in border regions: The Scandinavian case of Öresund, 1998–2001. *Annals of the Association of American Geographers*, *94*(4), 807-826.

Chasco, C., & Le Gallo, J. (2011, September). The impact of objective and subjective measures of air quality and noise on house prices: a multilevel approach for downtown Madrid. *In ERSA conference papers* (No. ersa11p168). European Regional Science Association.

Chen, F. M., Breiman, R. F., Farley, M., Plikaytis, B., Deaver, K., & Cetron, M. S. (1998). Geocoding and linking data from population-based surveillance and the US Census to evaluate the impact of median household income on the epidemiology of invasive Streptococcus pneumoniae infections. *American Journal of Epidemiology*, *148*(12), 1212-1218.

Chou, Y. H. (1995). Automatic bus routing and passenger geocoding with a geographic information system. In *Vehicle Navigation and Information Systems Conference, 1995. Proceedings. In conjunction with the Pacific Rim TransTech Conference. 6th International VNIS.'A Ride into the Future'* (pp. 352-359). IEEE.

Continelli, T., McGinnis, S., & Holmes, T. (2010). The effect of local primary care physician supply on the utilization of preventive health services in the United States. *Health & place*, *16*(5), 942-951.

Costello, S., Cockburn, M., Bronstein, J., Zhang, X., & Ritz, B. (2009). Parkinson's disease and residential exposure to maneb and paraquat from agricultural applications in the central valley of California. *American Journal of Epidemiology*, *169*(8), 919-926.

Davis, C. A., Fonseca, F. T., & Borges, K. A. (2003). A Flexible Addressing System for Approximate Geocoding. in V Brazilian Symposium on GeoInformatics (GeoInfo 2003), Campos do Jordão (SP),.

Dueker, K. J. (1974). Urban geocoding. *Annals of the Association of American Geographers*, 318-325.

ESRI, E. (1998). Shapefile technical description. Redland, CA: *An ESRI White Paper*.

Ge, X. (2005). U.S. Patent No. 6,934,634. Washington, DC: U.S. Patent and Trademark Office.

Gilboa, S. M., Mendola, P., Olshan, A. F., Harness, C., Loomis, D., Langlois, P. H., ... & Herring, A. H. (2006). Comparison of residential geocoding methods in population-based study of air quality and birth defects. *Environmental Research*, *101*(2), 256-262.

Goldberg, D. W. (2008). A geocoding best practices guide. Springfield, IL*: North American Association of Central Cancer Registries.*

Goldberg, D. W., & Cockburn, M. G. (2010). Improving geocode accuracy with candidate selection criteria. *Transactions in GIS*, *14*(s1), 149-176.

Goldberg, D. W., & Cockburn, M. G. (2012). The effect of administrative boundaries and geocoding error on cancer rates in California. *Spatial and Spatio-temporal Epidemiology*, *3*(1), 39-54.

Goldberg, D. W., Wilson, J. P., & Knoblock, C. A. (2007). From text to geographic coordinates: the current state of geocoding. *URISA Journal*, *19*(1), 33-46.

Goldberg, D. W., Wilson, J. P., Knoblock, C. A., Ritz, B., & Cockburn, M. G. (2008). An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*, *7*(1), 60.

Grubesic, T. H., & Matisziw, T. C. (2006). On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *International Journal of Health Geographics*, *5*(1), 58.

Gruenewald, P. J., & Remer, L. (2006). Changes in outlet densities affect violence rates. *Alcoholism: Clinical and Experimental Research*, *30*(7), 1184-1193.

Guo, F., Wang, X., & Abdel-Aty, M. A. (2010). Modeling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention*,*42*(1), 84-92.

Han, T., Alexander, M., Niggebrugge, A., Hollands, G. J., & Marteau, T. M. (2014). Impact of tobacco outlet density and proximity on smoking cessation: A longitudinal observational study in two English cities. *Health & Place*, 27, 45-50.

Here (2013) *Here. Map for life.* Retrieved from: http://here.com/navteq-redirect/.

Horner, M. W., Zook, B., & Downs, J. A. (2012). Where were you? Development of a time-geographic approach for activity destination re-construction. *Computers, Environment and Urban Systems*, *36*(6), 488-499.

Jaro, M. (1984). Record linkage research and the calibration of record linkage algorithms. In *Statistical Research Division Report Series SRD Report No. Census/SRD/RR-84/27*.

Kim, K., & Lahr, M. L. (2013). The impact of Hudson-Bergen Light Rail on residential property appreciation. *Papers in Regional Science*.

Krieger, N., Chen, J. T., Waterman, P. D., Rehkopf, D. H., Yin, R., & Coull, B. A. (2006). Race/ethnicity and changing US socioeconomic gradients in breast cancer incidence: California and Massachusetts, 1978–2002 (United States). *Cancer Causes & Control*, *17*(2), 217-226.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? the Public Health Disparities Geocoding Project. *American Journal of Epidemiology*, *156*(5), 471-482.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M. J., Subramanian, S. V., & Carson, R. (2003). Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The Public Health Disparities Geocoding Project (US). *Journal of Epidemiology and Community Health*, *57*(3), 186-199.

Krieger, N., Waterman, P., Chen, J. T., Soobader, M. J., Subramanian, S. V., & Carson, R. (2002). Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census-defined geographic areas-the Public Health Disparities Geocoding Project. *American Journal of Public Health*, *92*(7), 1100-1102.

Krieger, N., Waterman, P., Lemieux, K., Zierler, S., & Hogan, J. W. (2001). On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, *91*(7), 1114.

Los Angeles Almanac (2004) *City of Los Angeles & communities* Retrieved from: http://www.caltenantlaw.com/LAcommunities.htm

Lee, S. W. (2002). *U.S. Patent No. 6,397,208*. Washington, DC: U.S. Patent and Trademark Office.

Levine, N., & Kim, K. E. (1998). The location of motor vehicle crashes in Honolulu: a methodology for geocoding intersections. *Computers, Environment and Urban Systems*, *22*(6), 557-576.

Ludwig, I., Voss, A., & Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. In *Advancing Geoinformation Science for a Changing World* (pp. 65-84). Springer Berlin Heidelberg.

Mamalian, C. A., & La Vigne, N. G. (1999). *The use of computerized crime mapping by law enforcement: Survey results*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

Matthews, S. A., McCarthy, J. D., & Rafail, P. S. (2011). Using ZIP code business patterns data to measure alcohol outlet density. *Addictive Behaviors*, *36*(7), 777-780.

McElroy, J. A., Remington, P. L., Trentham-Dietz, A., Robert, S. A., & Newcomb, P. A. (2003). Geocoding addresses from a large population-based study: lessons learned. *Epidemiology*, *14*(4), 399-407.

Nicoara, G. (2005). Exploring the geocoding process: a municipal case study using crime data. Master's thesis, *Dallas, TX: University of Texas at Dallas*.

Ngamini Ngui, A., & Vanasse, A. (2012). Assessing spatial accessibility to mental health facilities in an urban environment. *Spatial and Spatio-temporal Epidemiology*, *3*(3), 195-203.

Oliver, M. N., Matthews, K. A., Siadaty, M., Hauck, F. R., & Pickle, L. W. (2005). Geographic bias related to geocoding in epidemiologic studies. *International Journal of Health Geographics*, *4*, 29.

O'Reagan, R. T. (1987). Geocoding theory and practice at the Bureau of the Census. Bureau of the Census.

Ozimek, A., & Miles, D. (2011). Stata utilities for geocoding and generating travel time and travel distance information. *Stata Journal*, *11*(1), 106.

Park, S. H., Bigham, J. M., Kho, S. Y., Kang, S., & Kim, D. K. (2011). Geocoding vehicle collisions on Korean expressways based on postmile referencing. *KSCE Journal of Civil Engineering*, *15*(8), 1435-1441.

Police Foundation, & United States of America. (2000). Geocoding in Law Enforcement: Final Report.

Qin, X., Parker, S., Liu, Y., Graettinger, A. J., & Forde, S. (2013). Intelligent geocoding system to locate traffic crashes. *Accident Analysis & Prevention*, *50*, 1034-1041.

Ratcliffe, J. H. (2001). On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. *International Journal of Geographical Information Science*, *15*(5), 473-485.

Ratcliffe, J. H. (2004). Geocoding crime and a first estimate of a minimum acceptable hit rate. *International Journal of Geographical Information Science*,*18*(1), 61-72.

Rischpater, R., & Au, C. (2013). Microsoft Mapping: Geospatial Development with Bing Maps and C#. New York, NY: Apress.

Robinson, J. C., Wyatt, S. B., Hickson, D., Gwinn, D., Faruque, F., Sims, M., ... & Taylor, H. A. (2010). Methods for retrospective geocoding in population studies: the Jackson Heart Study. *Journal of Urban Health*, *87*(1), 136-150.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (2006). Geocoding in cancer research: a review.*American Journal of Preventive Medicine*, *30*(2), S16-S24.

Rushton, G., Armstrong, M. P., Gittler, J., Greene, B. R., Pavlik, C. E., West, M. M., & Zimmerman, D. L. (Eds.). (2010). *Geocoding Health Data: The Use of Geographic codes in cancer prevention and control, research and* practice. Boca Raton, FL: CRC Press.

Schootman, M., Sterling, D. A., Struthers, J., Yan, Y., Laboube, T., Emo, B., & Higgs, G. (2007). Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology*, *17*(6), 464-470.

Snow, J. (1855). On the mode of communication of cholera. London: John Churchill.

Texas A&M GeoServices (2013) *TAMU GeoServices* Retrieved from: http://geoservices.tamu.edu/

Taranenko, A., & Taranenko, I. (2011). *U.S. Patent No. 8,015,196*. Washington, DC: U.S. Patent and Trademark Office.

U.S. Census Bureau TIGER/Line. U.S. Census Bureau 2013. Retrieved from: http://www.census.gov/geo/www/tiger.

Vieira, V. M., Howard, G. J., Gallagher, L. G., & Fletcher, T. (2010). Research Geocoding rural addresses in a community contaminated by PFOA: a comparison of methods.

Vine, M. F., Degnan, D., & Hanchette, C. (1997). Geographic information systems: their use in environmental epidemiologic research. *Environmental Health perspectives*, *105*(6), 598.

Wheeler, D. C., Ward, M. H., & Waller, L. A. (2012). Spatial-temporal analysis of cancer risk in epidemiologic studies with residential histories. *Annals of the Association of American Geographers*, *102*(5), 1049-1057.

Wieczorek, J., Guo, Q., & Hijmans, R. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International journal of geographical information science, 18(8), 745-767.

Yang, D. H., Bilaver, L. M., Hayes, O., & Goerge, R. (2004). Improving geocoding practices: evaluation of geocoding tools. *Journal of Medical Systems*, *28*(4), 361-370.

Zandbergen, P. A. (2008). A comparison of address point, parcel and street geocoding techniques. *Computers, Environment and Urban Systems*, *32*(3), 214-232.

Zarem, M., Vuillermet, E., & DeAguiar, J. (2006). *U.S. Patent Application 11/367,911*.

Zimmerman, D. L., & Li, J. (2010). The effects of local street network characteristics on the positional accuracy of automated geocoding for geographic health studies. *International Journal of Health Geographics*, *9*(10), 1-11.

Zhan, F. B., Brender, J. D., De Lima, I., Suarez, L., & Langlois, P. H. (2006). Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology*, *16*(11), 842-849.